A Prompt-independent and Interpretable Automated Essay Scoring Method for Chinese Second Language Writing

Yupei Wang
School of Science
Beijing Jiaotong University
yepiwong@gmail.com

Renfen Hu[✉*]
Institute of Chinese Information Processing
Beijing Normal University
irishu@mail.bnu.edu.cn

Abstract

With the increasing popularity of learning Chinese as a second language (L2), the development of an automatic essay scoring (AES) method specially for Chinese L2 essays has become an important task. To build a robust model that could easily adapt to prompt changes, we propose 90 linguistic features with consideration of both language complexity and correctness, and introduce the Ordinal Logistic Regression model that explicitly combines these linguistic features and low-level textual representations. Our model obtains a high QWK of 0.714, a low RMSE of 1.516 and a considerable Pearson correlation of 0.734. With a simple linear model, we further analyze the contribution of the linguistic features to score prediction, revealing the model's interpretability and its potential to give writing feedback to users. This work provides insights and establishes a solid baseline for Chinese L2 AES studies.

1 Introduction

Automatic Essay Scoring(AES) is one of the most important Natural Language Processing (NLP) applications in the field of education (Page, 1966; Ke and Ng, 2019), and has been widely used in standardized language tests (Burstein and Chodorow, 1999; Attali and Burstein, 2006). However, existing works mainly focus on the scoring of English essays (Yannakoudakis et al., 2011; Taghipour and Ng, 2016; Alikaniotis et al., 2016) or Chinese essays by native speakers and minority learners (Chang and Lee, 2009; Peng et al., 2010; Song et al., 2020). Although Chinese second language (L2) acquisition has enjoyed an increasing boom in recent decades, the AES system designed for Chinese L2 writing has received much less attention.

Meanwhile, existing AES methods face two important challenges. Firstly, the scoring models are mostly built in a prompt-dependent style, i.e. training and testing for each specific prompt. It requires to collect prompt-specific data, yielding great costs in dataset construction (Attali and Burstein, 2006). Besides, the built models are of weak generalization capabilities and cannot be used to score essays of other prompts. Secondly, although neural network methods have achieved great success in NLP tasks, the gains in neural AES systems are far from being satisfactory. For example, Mayfield and Black (2020) find that fine-tuning BERT produces similar performance to classical models at significant additional cost. Apart from the costs, the deep neural models are also weak in interpretability of the results. However, it is a very important property for AES users who expect to get feedback on the writing (Woods et al., 2017; Ke et al., 2018; Ke and Ng, 2019), not just a score.

To solve the above problems, this paper proposes a prompt-independent and interpretable AES method for Chinese L2 writing. Specifically, we build prompt-independent models that could make full use of L2 writing data, and make predictions without the prompt limitations. For interpretability considerations, we extract 90 linguistic indices on accounting of the usage of characters, words, clauses, collocations, dependency structures, syntactic constructions which are emphasized in Chinese L2 acquisition, and 5 indices that address different types of writing errors. Furthermore, we integrate these linguistic and

*Corresponding author.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, August 13 - 15, 2021.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

correctness indices into text representations, and introduce the Ordinal Logistic Regression (OLR) model to the AES task for Chinese second language writing. Our model achieves a high quadratic weighted Kappa (QWK) score of 0.714, a low Root Mean Square Error (RMSE) of 1.516, and a high Pearson coefficient of 0.734, performing much better than the classical machine learning models and neural network baselines.

The contribution of this paper is two-fold. (1) Instead of building prompt-specific essay scoring models, it presents a generic model that could make full use of writing data, and score general narrative and argumentative essays. (2) By integrating various dimensions of linguistic features which are emphasized in Chinese L2 acquisition, the models are both effective and interpretable when making predictions. The source code of our method is publicly available[0].

2 Related Work

2.1 Prompt-specific vs. Prompt-independent

Most existing AES methods are built as a prompt-specific style, i.e. training and testing with prompt-specific data (Taghipour and Ng, 2016; Dong et al., 2017; Woods et al., 2017) or relying on prompt-specific features (Attali and Burstein, 2006). They can sometimes achieve better results than those trained regardless of prompts. However, the application of the models are limited to specific topics and they could not make full use of the data. In addition, it will be costly and time-consuming to obtain training data each time when a new prompt is introduced. Two approaches have been developed to improve the situation. One is to directly use all the data (Alikaniotis et al., 2016), ignoring the differences of the prompts. Another method is domain adaptation (Phandi et al., 2015; Dong et al., 2017; Cao et al., 2020), which could make better use of available essays in all prompts and make the model robust to the change of prompts.

2.2 Interpretability and Feedback

Many success have been achieved by holistic scoring of essays. However, this method faces challenges in providing effective feedback to the students due to its poor interpretability, especially for those neural models (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Dong et al., 2017). Taghipour and Ng (2016) examine the score variations for three essays after processing each word by the neural network, and find that the model is able to learn essay length and essay content. Alikaniotis et al. (2016) visualize the "quality" of the word vectors. However, these methods could only give very shallow explanations of the model behaviors, and are not able to give end user feedback.

A mainstream approach to solve this problem is to score the essays from different dimensions, such as coherence, argument strength, prompt adherence and organization (Persing et al., 2010; Persing and Ng, 2013; Persing and Ng, 2014; Persing and Ng, 2015). Although it can help the students to understand the shortcomings of their essays, more detailed feedback is still welcome. Woods et al. (2017) developed a model-driven sentence selection approach, which can give students sentence-level advice in detail. Ke et al. (2018) identify a set of attributes that can explain an argument's persuasiveness and annotate each argument in corpus with the values of these attributes.

2.3 Automatic Essay Scoring of Chinese Essays

The research on Chinese AES has receieved much less attention compared with English AES, and lots of work focus on essays of native Chinese speakers. However, the characteristics of L2 essays are quite different from those by native speakers (Cao and Deng, 2012; Wu et al., 2019). A series of works show that linguistic complexity features are quite effective in measuring the quality of Chinese L2 writing (Huang et al., 2014; Wang, 2017; Wu, 2018; Wu and Xing, 2020). However, most of these studies only examine their method with a small set of essays of limited prompts. The effectiveness of these features on large-scale datasets remain to be discussed, and their roles in AES systems are also worthy of further exploration. Motivated by previous works, this paper proposes a prompt-independent AES approach

[0]https://github.com/iris2hu/L2C-rater

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

for Chinese L2 writing, which integrates a wide scope of linguistic complexity features to enhance the interpretability of the models.

3 The Proposed Method

3.1 The Interpretable Representations of Essay Features

We extracted three types of interpretable features to represent the L2 essays, including linguistic complexity, writing errors and various dimensions of textual features. We measure the linguistic complexity with consideration of the diversity and sophistication of characters, words, clauses, collocations, dependency structures and syntactic constructions. Regarding the writing errors, we build punctuation, character, vocabulary, sentence and discourse level indices. In terms of textual features, we introduce characters, words, ngrams and part-of-speeches.

3.1.1 Linguistic Complexity Features

The effectiveness of linguistic complexity features has been well addressed in predicting the Chinese L2 writing quality (Huang et al., 2014; Wang, 2017; Wu et al., 2019). In addition, they could provide direct feedback to the users on accounting of the usages of different linguistic units, which is highly explainable. Therefore, this paper designs and constructs a comprehensive set of linguistic complexity measures of Chinese L2 writing and integrate them into the representations of L2 essays.

It should be noted that when designing the feature set, it is not applicable to directly transfer the ones that work in English AES or AES systems for Chinese native speakers to Chinese L2 AES, because Chinese has a lot of language-specific features that are emphasized in second language acquisition. Hu (2021) pointed out that indices based on language-specific features have stronger predictive power and higher efficiency in predicting the L2 writing scores. Hence the linguistic complexity feature set for Chinese L2 AES should take into account both the language-independent and language-specific features. In this paper, we build 90 linguistic indices of writing quality from the following dimensions. A full list of the indices and their descriptions can be seen in the Appendix A.

Chinese characters and vocabulary. We build four indices in this dimension, including the number of Chinese characters, the number of Chinese words, lexical diversity and lexical sophistication. The lexical diversity index is computed as the root type token ratio (RTTR) of words. The lexical sophistication is built as the ratio of sophisticated words. In this study, we identify the words of HSK-5 level, HSK-6 level and out of the HSK vocabulary as the sophisticated words.

Sentences and clauses. Seven indices are proposed to measure the sentence and clausal complexity, including the mean length of sentences, the mean length of clauses, the mean length of T-units, number of clauses per sentence, number of T-units per sentence, the mean depth of the dependency trees and the max depth of the the dependency trees.

Collocations and bigrams. We introduce 21 collocation-based indices and two bigram-based indices in this dimension. First, eight types of collocations are considered by following Hu and Xiao (2019)'s work, including Verb-Object (VO), Subject-Predicate (SP), Adjective-Noun (AN), Adverb-Predicate (AP), Classifier-Noun (CN), Preposition-Postposition (PP), Preposition-Verb (PV) and Predicate-Complement (PC), where the former four (VO, SP, AN, AP) are universal collocation types that exist in different languages, while the later four (CN, PP, PV, PC) are language-specific types that have been greatly emphasized in Chinese second language acquisition. Similar to lexical diversity, the collocation diversity is built as the RTTRs of different types of collocations, including all the collocations, language-specific collocations, language-independent collocations, and each type of the collocations, resulting in 11 diversity indices. Besides, to measure the collocation sophistication, we introduce the ratio of low frequency collocations and language-specific collocations by following Hu (2021)'s work[1]. Also, the ratio of each type of collocations is computed. To cover more language usages, we implement the bigram diversity and sophistication as well by considering the bigrams as a specific type of collocations.

Dependency structures. The eight types of collocations (Hu and Xiao, 2019) are extracted from dependency parsing trees with rule-based methods. Although they can well reflect the important knowl-

[1]https://github.com/iris2hu/Chinese-collocation-complexity

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

edge in Chinese L2 acquisition, there are still two problems. One is that they only target at a part of the syntactic relations, hence lacking a whole picture of the syntactic structures. Another is that the collocation diversity and sophistication are not able to measure the fine-grained phrasal complexity underlying the structures, e.g. the number and length of the modifiers. To address the above two questions, this paper proposes 41 dependency based indices that measure the distance, diversity and ratio of all the dependency triples. In this work, we use the LTP dependency parser[2] and 13 dependency relations are considered when building the corresponding indices.

Constructions. The acquisition of grammatical constructions is one of the most important aspects of Chinese L2 teaching and learning (Lu, 2000; Sun, 2016; Zhao, 2018). Both the standardized language test developers and textbook editors make great efforts in designing appropriate construction lists for certain levels of learners.

Consider the importance of construction knowledge in Chinese L2 acquisition, this paper proposes to measure the density and ratio of constructions with regarding to their levels. Specifically, we employ the construction list from the General Syllabus of International Chinese Teaching (Hanban, 2009) which include 62 constructions of five levels. After automatic recognition of the constructions, we build 15 indices to reflect the density and ratio of different levels of constructions.

3.1.2 Writing Error Features

In addition to the linguistic complexity, the correctness of L2 production also plays an important role in automatic essay scoring or speech rating. Therefore, we adopt five indices of writting errors, i.e. the number of punctuation errors, Chinese character errors, word level errors, sentence level errors and discourse level errors with reference to the annotation in HSK Dynamic Composition Corpus[3].

3.1.3 Multi-granularity Text Features

The high correlation between lexical complexity and writing scores have been witnessed in many studies (Peng et al., 2010; Wang, 2017). However, it is still beneficial to further retain the full picture of the textual features.

To represent a text, we extract character, word and part-of-speech unigrams, bigrams and trigrams as features since they could reflect multi-granularity language usages, and could be more explainable than neural representations e.g. word embeddings. We use the tf–idf weighted representations of these features, and each essay can be represented as a text vector:

$$TextVec = (tfidf_1, tfidf_2, \ldots, tfidf_N) \quad (1)$$

Where N denotes the total number of unique language units that appear in the corpora.

For the above linguistic complexity, writing error and multi-granularity text features, we conduct preliminary experiments to make feature selection and combination. The detailed process will be introduced in the Experiment section.

3.2 The Ordinal Logistic Regression Model

Automatic essay scoring is mainly built as classification or regression tasks. Although the classification models can achieve good results, they treat each score as an independent category, hence losing the ordering information. While for linear regression, it may suffer from violations of modeling assumptions because of the small, discrete, range of possible scores (Woods et al., 2017).

To address the above questions, this paper proposes to use the Ordinal Logistic Regression (OLR) model in Chinese L2 AES since the OLR method is an effective classification method for ordinal categories (Rennie, 2005). In the classification problem with ordinal classes, the loss of mis-predicting a certain category into different categories should not be the same, e.g. predicting 2 as 3 vs. predicting 2 as 6. The traditional classification loss functions need to be improved to adapt to this relationship (Rennie and Srebro, 2005). Woods et al. (2017) firstly introduce this method into English AES study and achieve

[2]https://github.com/HIT-SCIR/ltp

[3]http://hsk.blcu.edu.cn/

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

impressive results. Inspired by their work, this paper introduces the OLR models into Chinese L2 AES and compare its effectiveness to multiple classical machine learning and neural baselines.

A practical loss of ordinal classification is threshold-based, which is specifically divided into Immediate-threshold loss and All-threshold loss. The former is is actually a degradation of the latter, which is what we actually use. All-threshold loss are represented as (2):

$$\mathrm{Loss}_{\mathrm{AT}}(z) = \sum_{k=1}^{l-1} f\left(s(k;i)\left(\theta_k - z\right)\right) \quad s(k;i) = \begin{cases} -1 & k < i \\ +1 & k \geq i \end{cases} \tag{2}$$

where z is a specific predicted value, and (θ_{i-1}, θ_i) refers to the "correct" segment, and $f(\cdot)$ could be any kind of loss function for multiclass classification problem.

In this study, we employ a very large feature space, requiring regularization to alleviate possible overfitting. Thus the Regularized Logistic Regression (RLR) minimization objective is defined as

$$\mathrm{Loss}_{\mathrm{RLR}} = \sum_{i=1}^{N} \log\left(1 + \exp\left(-y_i \cdot \mathbf{x}_i^T \mathbf{w}\right)\right) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{3}$$

Defining $h(z) := \log(1 + \exp(z))$ and bringing $Loss_{\mathrm{RLR}}(\cdot)$ into $Loss_{AT}(\cdot)$ as $f(\cdot)$, we have the minimization objective for the All-threshold version of Ordinal Logistic Regression(OLR-AT)

$$\mathrm{Loss}_{\mathrm{ATL}} = \sum_{i=1}^{N}\left[\sum_{k=1}^{y_i-1} h\left(\theta_k - \mathbf{x}_i^T\mathbf{w}\right) + \sum_{k=y_i}^{l-1} h\left(\mathbf{x}_i^T\mathbf{w} - \theta_k\right)\right] + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{4}$$

where label $k \in \{1, \ldots, l\}$ corresponds to the segment (θ_{k-1}, θ_k). θ_0 and θ_l denotes $-\infty$ and $+\infty$ respectively. $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d$ are training examples while $\{y_1, \ldots, y_n\}$, $y_i \in \{1, \ldots, l\}$ are their labels.

4 Experiments

4.1 Dataset and Preprocessing

In the experiments, we use the essay data from HSK Dynamic Composition Corpus. HSK is a standardized test of Chinese language proficiency for non-native Chinese speakers. The essays are rated from 40 points to 95 points with five as an interval, yielding 12 different categories. The mean score is 69.499 and the Standard Deviation is 10.980. We use the 10277 argumentative and narrative essays (over 3.7 million Chinese characters) from the corpus to train and test the AES model. For a reliable evaluation, we conduct 5-fold cross validation. First, 1477 essays are randomly selected as the test set, and the remaining 8800 essays are split into five groups. Each time four groups are used for training and the left one is used as the development set, which helps to find the optimal parameters. Therefore, all the experiments are conducted five times and the average results on test set is reported.

As the essays in the corpus are manually labeled with different types of writing errors. After retrieving the writing error indices, we carefully remove the annotation tags and transform the essays to their original states, i.e. the original texts written by the test takers. Then we use the method proposed in Section 3.1.1 to obtain the 90 linguistic complexity indices. For multi-granularity text representations, we use `jieba` to conduct word segmentation and POS tagging, and the TfidfTransformer in `scikit-learn` to get the feature weights.

4.2 Feature Selection

To examine the predictive power of different types of linguistic complexity and writing error indices, we conduct step-wise linear regression in each dimension, and the result can be seen in Table 1. It suggests that the all of the six dimensions of indices could explain the score variances to some extent, where the indices built upon Chinese characters and vocabulary, collocations and bigrams, and dependency structures have stronger predictive power than the indices in other dimensions.

Dimension	R	R^2
Chinese characters and vocabulary (4, 3)	0.648	0.420
Sentences and clauses (7, 4)	0.197	0.039
Collocations and bigrams (23, 8)	0.587	0.345
Dependency structures (41, 16)	0.610	0.372
Constructions (15, 9)	0.248	0.061
Writing Error Features (5, 4)	0.254	0.065

Table 1: Step-wise regression results in each dimension. The numbers in brackets denote the number of indices entered and remained in the step-wise regression respectively.

Before building the essay scoring model, we make feature selection of the proposed linguistic indices to avoid multicollinearity problem. For the 90 linguistic complexity indices, we select 33 indices with the step-wise regression method. After integrating the five writing error features, the step-wise regression model yields 31 effective features. In the following experiments, these two feature sets are used as the `ling` and `ling+err` settings. The selected features can be seen in Appendix A. For the multi-granularity textual features, we examine different feature combinations in preliminary experiments and find that the combination of word unigrams and pos features could achieve optimal performance with efficient feature space, thus they are used as the `text` setting.

4.3 Models, Parameters and Evaluation Metrics

For the text representations, the min term frequency is set to 10. For OLR-AT model, the penalty coefficient λ is set to 1.0. To make a comparison, we build two types of baselines in the experiments, including regression-based and tree-based machine learning models that use the same input features as our OLR method, and an effective neural AES model introduced by Taghipour and Ng (2016) which extract features automatically and implicitly.

Linear Regression. Linear Regression (LiR) refers to the process of fitting a multi-dimensional linear function to all data points as much as possible. Adding the L1 or the L2 regular term to the cost function yields two variants, i.e. LASSO and Ridge Regression.

Logistic Regression. Logistic Regression(LoR) is a generalized linear model for binary classification. In multi-class scenario, it can be implemented with a One vs. Rest scheme. In our experiment, we set the maximum iteration threshold to a large value (1000) to ensure that the algorithm converges as much as possible.

Random Forest Regression. Random Forest (RF) is based on bagging mechanism and contains multiple decision trees generated in parallel. Each decision tree randomly selects a part of the feature vector for training, and the output is the average results of the trees. In the experiments, the maximum tree depth of the Random Forest Regression is set to 40.

XGBoost Regression. XGBoost is an improved version of the gradient boosting algorithm GDBT. In the experiments, the maximum tree depth is constrained to 3, the number of estimators is constrained to 300, the learning rate is set to 0.05, and the gamma is set to 5.

CNN+LSTM. The CNN+LSTM architecture is a classical neural baseline for English AES task (Taghipour and Ng, 2016). Here we introduce this model to Chinese L2 AES. We use 300-dim Chinese word vectors[4] pre-trained on Sogou news corpus. We train the network for 20 epochs and the batch size is 32. The training is stopped when the model does not make further improvement after 1000 batches of training. The vocabulary size is set as 20000. Other settings align with Taghipour and Ng (2016).

Att-BLSTM. The Att-BLSTM architecture was first proposed for relation classification task (Zhou et al., 2016). Here we adjust its output from a vector to a scalar so that it can be used for regression task. For other settings, e.g. the use of word embeddings, epochs and batch size, are consistent with those in CNN+LSTM. The model parameters align with Zhou et al. (2016).

[4]https://github.com/Embedding/Chinese-Word-Vectors

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, August 13 - 15, 2021.

There are many metrics (Yannakoudakis and Cummins, 2015) that can measure the correlation and consistency between the outputs of the AES system and the scores of human experts. In this work we employ three of them: Quadratic Weighted Kappa(QWK), Root Mean Square Error(RMSE) and Pearson coefficient(Pears.). QWK is widely adopted for evaluating AES methods (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Woods et al., 2017; Song et al., 2020), RMSE is a standard way to measure the error of models, while Pearson coefficient could reflect scoring consistency.

4.4 Results

In the experiments, the machine learning methods use four different feature sets as described above: `ling`, `ling+err`, `ling+text` and `ling+err+text`. When using the combination of linguistic and text features, we concatenate the feature matrices. The CNN+LSTM and Att-BLSTM baselines employ two settings by initializing the word vectors randomly or with the pre-trained Sogou embeddings. Table 2 shows the results of our OLR-AT model and other baselines.

Method	Mode	QWK	RMSE	Pears.	Mode	QWK	RMSE	Pears.
LiR	ling	0.640	**1.636**	**0.679**	ling+text	0.269	3.576	0.299
	ling+err	**0.668**	**1.585**	**0.702**	ling+err+text	0.276	3.557	0.307
LoR	ling	0.598	1.813	0.620	ling+text	0.641	1.720	0.663
	ling+err	0.640	1.715	0.661	ling+err+text	0.663	1.667	0.681
RFR	ling	0.625	1.657	0.668	ling+text	0.652	1.603	0.694
	ling+err	0.655	1.601	0.695	ling+err+text	0.667	1.575	0.706
XGBR	ling	0.576	1.690	0.652	ling+text	0.587	1.676	0.659
	ling+err	0.613	1.625	0.687	ling+err+text	0.621	1.616	0.690
CNN+LSTM	Random	0.496	1.845	0.551	Sogou	0.504	1.831	0.560
Att-BLSTM	Random	0.520	1.825	0.568	Sogou	0.531	1.812	0.578
OLR-AT	ling	**0.644**	1.650	0.674	ling+text	**0.697**	**1.554**	**0.718**
	ling+err	0.666	1.616	0.691	ling+err+text	**0.714**	**1.516**	**0.734**

Table 2: Results of Chinese L2 AES. The **bold** denotes the best result under the same feature setting.

It can be seen that the OLR-AT model on `ling+err+text` feature setting achieves the best performance overall, suggesting the effectiveness of the OLR model and the use of feature combinations. The different models and feature settings also yield different results. We make comparisons of them as below.

Feature settings. The OLR-AT and machine learning methods all integrate four feature settings. First, `ling+err` brings consistent improvements to `ling` after integrating the error information. This echoes the emphasis on writing error information in HSK standards (Dan, 2009). Second, except for Linear Regression (LiR), all models obtain the best results under `ling+err+text`. It is worth noting that the very simple Linear Regression model achieves almost the best results under `ling` and `ling+err`. It indicated that LiR, as a simple, effective and interpretable model, might be weak in dealing with high-dimension feature space. To solve this problem, we could introduce parameter regularizations, which will be further explored in the Discussion section.

Models. From a model point of view, we firstly notice that the neural baseline CNN+LSTM does not achieve comparable results of OLR-AT and other machine learning methods, suggesting that it is not applicable to directly transfer the method that work in English to Chinese[5]. Similarly, the neural model Att-BLSTM, which performs well in other tasks such as relation classification, is also difficult to obtain competitive results. It is worth noting that the OLR-AT model surpasses almost all other models. Also, after adding text features to `ling+err`, the performance of OLR-AT improves by 7.2%, compared with 3.6% of Logistic Regression, 1.8% of Random Forest and 1.3% of XGBoost.

Since the HSK dataset does not release scores of different human raters, we are not able to compare the Chinese AES results to human performance. As a reference, the English ASAP (Automated Student

[5]As a reference, the CNN+LSTM model achieves a high QWK on the English ASAP dataset: 0.717 (AES and Rater1) and 0.710 (AES and Rater2). The QWK of two human raters is 0.754 (Taghipour and Ng, 2016).

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, August 13 - 15, 2021.

Assessment Prize) dataset reported the average between-rater QWK as 0.754[6]. Given our best model (OLR-AT under `ling+err+text`) achieves a QWK of 0.714, it indicates that our method could be a solid work for Chinese L2 AES task. Next, we will make further discussion of the models' errors and shed some light on future work. In addition, we will explore the improved Linear Regression model since it is a simple yet the most explainable model which has the potential to offer users feedback.

5 Discussion

5.1 Analysis on Confusion Matrix

To illustrate the models' behaviors, Figure 1 shows the confusion matrix of the OLR-AT model under `ling+err+text`.

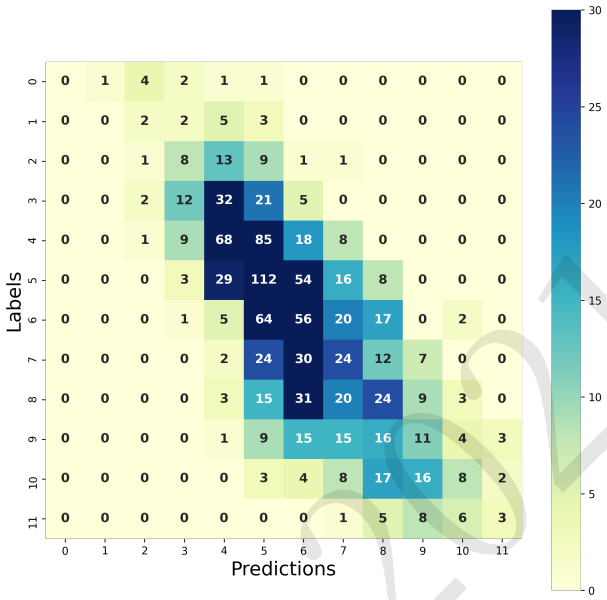

Figure 1: Confusion Matrix of OLR-AT Results

It can be seen that the darker blocks distributed around the main diagonal, indicating a high QWK of the model. However, there are still some outliers deviating from the main diagonal. We manually checked the essays that are scored too high or too low (± 2 classes), and find the errors are mainly due to the following reasons:

- For essays with high predicted scores, they typically have a high proficiency of language uses, but the contents deviate from their prompts, or (for argumentative essays) are lack of organization when expressing opinions. Existing feature sets and algorithms cannot detect these writing flaws.
- For essays with low predicted scores, we find some rating exceptions by the human raters, e.g. giving high scores to unfinished essays. Since the length of the essay is an important feature in our model, the AES scores lower than the human raters. Therefore, on the whole, the algorithm's ability to capture current features is in place, and the evaluation effect is relatively reliable.

From the above analysis, we find that it would be helpful to further introduce prompt-essay relevance measures, as well as the discourse level indices e.g. cohesion and coherence. We will conduct the research from these aspects in future work.

5.2 Revisiting Linear Regression: Interpretability and Potential of Providing Feedback

Why does the result of Linear Regression drop so significantly after introducing the `text` representations? We speculate that the reason lies in the high-dimension and sparse feature space of text representations, which could easily lead to over-fitting of linear model. To verify this, we implement Linear

[6]ASAP is the most popular dataset in English AES studies, and the dataset can be downloaded at Kaggle.

Regression under `text` setting, and compare it with the results of `ling+text` and `ling+err+text` as shown in Table 3. It can be easily seen that the `text` only setting has a low performance, and integrating text features to linguistic features does not make improvements to `ling` and `ling+err` settings.

Mode	QWK	RMSE	Pears.
text	0.207	3.787	0.232
ling+text	0.269	3.576	0.299
ling+err+text	0.276	3.557	0.307

Table 3: The results of Linear Regression with different feature sets.

Further, we use one of the variants of linear regression - Ridge Regression, that is, adding an L2 regular term to the objective function of Linear Regression. Ridge regression loses unbiasedness in exchange for high numerical stability(Hoerl and Kennard, 1970). The results are shown in Table 4. Ridge Regression greatly alleviated the over-fitting phenomenon. Different from Linear Regression, Ridge makes clear improvements after integrating text features, surpassing the tree-based methods even. It is close to OLR-AT in terms of QWK, and even scores slightly better than OLR-AT on RMSE and Pearson correlation coefficient. The reason is that the OLR-AT method focuses on correct classification, while Ridge aims to minimize the sum of squares of deviations under the constraint of regular terms.

Method	Mode	QWK	RMSE	Pears.	Mode	QWK	RMSE	Pears.
LiR	ling	0.640	1.636	0.679	ling+text	0.269	3.576	0.299
	ling+err	0.668	1.585	0.702	ling+err+text	0.276	3.557	0.307
Ridge	ling	0.636	1.640	0.676	ling+text	0.694	1.538	0.723
	ling+err	0.667	1.585	0.702	ling+err+text	0.709	1.510	0.735

Table 4: The comparison of Linear Regression and Ridge Regression

The power of Ridge Regression, a simple linear model, inspires us to explore the interpretation the results and the possibility to provide feedback to L2 students. Note that Linear Regression under `ling+err` setting achieves better performance than that of OLR-AT, indicating that the linear relationship dose exist between low-dimension linguistic features and the writing scores, and the linguistic features explain a large amount of score variances. Thus, studying how linear model uses linguistic features to score essay helps understand what an excellent essay should be like in a model perspective. Figure 2 shows 31 box plots of the selected `ling+err` features, showing the effect of each feature on essay scores. In addition, we choose three essays of high score (95 points), medium score (65 points), and low score (45 points) respectively and mark their effect value in the box plot to show how each feature contributes to their final scores. The three example essays can be seen in Appendix B. From the effects in Figure 2, we can clearly see the pros and cons of each essay, and provide corresponding feedback as below:

- *Essay of high score (the green triangle).* The student can write a long essay and use diverse and sophisticated words. In terms of syntactic usages, the essay employs sufficient syntactic constructions, but the use of language-specific structures is limited. Considering the correctness, the student can write Chinese characters with a high accuracy, but there are still some word, sentence and punctuation errors.

- *Essay of medium score (the blue circle).* The essay is of medium length and lexical diversity. The vocabulary used in the essay is relatively simple. From the syntactic view, the student is able to produce language-specific structures and idiomatic expressions skillfully. The student can use most words correctly, but the essay still contains some character, sentence and discourse errors.

- *Essay of low score (the red cross).* The essay is short with a limited word and collocation vocabulary, but the student is able to use some sophisticated words and elementary level constructions. Also, The student can produce correct and fluent text with relatively few mistakes.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

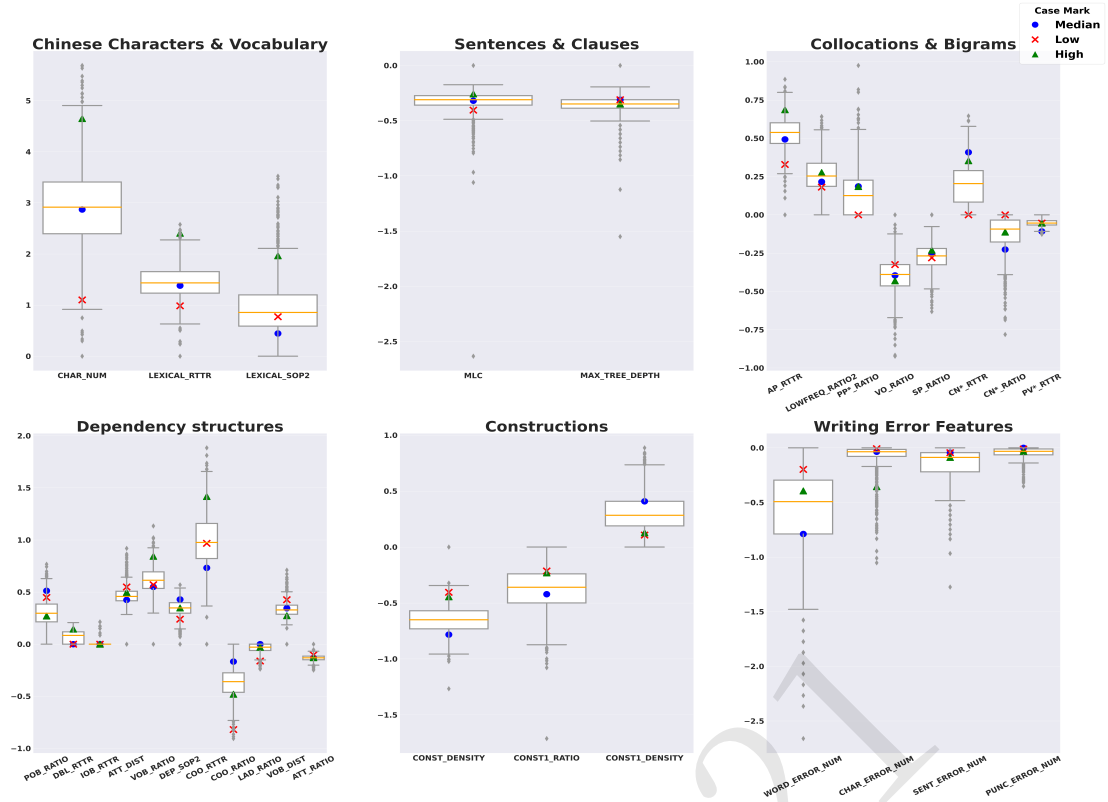

Figure 2: Effect plots of 31 selected features in `ling+err` setting. We use product of linear model coefficients and feature values $\omega_i x_i$ to convey the effect, where ω_i denotes the coefficient of feature i from the linear model, and x_i is the corresponding feature value.

6 Conclusion and Future Work

In this paper, we propose a prompt-independent and interpretable AES method for Chinese L2 writing. We build explainable representations of both the linguistic and text features, and the threshold-based Ordinal Logistic Regression model is introduced to our Chinese L2 AES task. The result on OLR-AT model under `ling+err+text` setting obtains a high QWK score of 0.714, a low RMSE of 1.516, and a high Pearson coefficient of 0.734. Further, we find that with our method and the feature set, the model is explainable and has the potential to offer users feedback on the writing. This work provides insights and a solid baseline for AES studies of Chinese L2 writing.

At present, our method has integrated linguistic complexity, writing correctness and text features of essays. It still needs further study on developing prompt-essay relevance measures, as well as the discourse level indices e.g. cohesion and coherence. Just as noteworthy, in Table 2, the CNN+LSTM architecture seems to perform poorly, far inferior to its excellent performance on the English ASAP dataset. This explains to a certain extent how different the scoring standards for Chinese L2 essays and those for native English speakers. It should be pointed out that we are not denying the possibility of applying neural networks to Chinese L2 AES task. Since CNN+LSTM structure itself is relatively simple, it cannot fully detect the features that a Chinese L2 AES task needs. As a model that has not been specially adjusted for the Chinese AES task, Att-BLSTM's improvement over CNN+LSTM has shown that neural networks have potential to achieve better results. Thus in future work, neural networks with stronger learning abilities, together with a good interpretation method may play important roles in this task.

Acknowledgements

This research was supported by the National Social Science Fund of China (No. 18CYY029), the National Natural Science Fund of China (No. 62006021), and the National Training Program of Innovation and Entrepreneurship for Undergraduates (No. 202110004069).

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725.

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

Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative english speakers. In *Computer mediated language assessment and evaluation in natural language processing*.

Xianwen Cao and Sujuan Deng. 2012. The contrastive analysis of the writing performance in chinese as l1 and l2:based on the chinese compositions on the same topic among chinese senior elementary students and vietnam senior university students. *TCSOL Studies (huáwén jiāoxué yǔ yánjiū)*, (2):39–46.

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

Tao-Hsing Chang and Chia-Hoang Lee. 2009. Automatic chinese essay scoring using connections between concepts in paragraphs. In *2009 International Conference on Asian Language Processing*, pages 265–268. IEEE.

Nie Dan. 2009. A historical account of the assessment criteria for hsk writing [j]. *Journal of Yunnan Normal University (Teaching and Research on Chinese as a Foreign Language Edition)(Yúnán shīfàn dàxué xuébào(duìwài hànyǔ jiàoxué yǔ yánjīu bǎn))*, 7(6).

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.

Hanban. 2009. *General Course Syllabus for International Chinese Language Teaching*. Beijing: Foreign Language Teaching and Research Press.

Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Renfen Hu and Hang Xiao. 2019. The construction of chinese collocation knowledge bases and their application in second language acquisition. *Applied Linguistics (yǔyán wénzì yìngyòng)*, (1):135–144.

Renfen Hu. 2021. On the relationship between collocation-based syntactic complexity and chinese second language writing. *Applied Linguistics (yǔyán wénzì yìngyòng)*, (1):132–144.

Zhi'e Huang, Jiali Xie, and Endong Xun. 2014. Study of feature selection in hsk automated essay scoring. *Computer Engineering Applications(Jìsuànjī gōngchéng yǔ yìngyòng)*, (6):118–122+126.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, pages 6300–6308.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *IJCAI*, pages 4130–4136.

Jianming Lu. 2000. The grammatical teaching in chinese second language teaching. *Language Teaching and Research (yǔyán jiāoxué yǔ yánjiū)*, (3):1–8.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.

E. B. Page. 1966. Grading essays by computer: Progress report. *Proceedings of the Invitational Conference on Testing Problems*, page 87–100.

Xingyuan Peng, Dengfeng Ke, Zhenbiao Chen, and Bo Xu. 2010. Automated chinese essay scoring using vector space models. In *2010 4th International Universal Communication Symposium*, pages 149–153. IEEE.

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.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

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.

Isaac Persing and Vincent Ng. 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)*, pages 543–552.

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.

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.

Jason D.M. Rennie and Nathan Srebro. 2005. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*, volume 1. Citeseer.

Jason D.M. Rennie. 2005. Ordinal logistic regression. http://people.csail.mit.edu/jrennie/writing/olr.pdf.

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.

Dejin Sun. 2016. Two theoretical issues on the studies of the pedagogic grammar of tcsl. *Language Teaching and Research (yǔyán jiāoxué yǔ yánjiū)*, (2):30–39.

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.

Yixuan Wang. 2017. The correlation between lexical richness and writing score of csl learner——the multi-variable linear regression model and equation of writing quality. *Applied Linguistics(yǔyán wénzì yìngyòng)*, (2):93–101.

Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. 2017. Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2071–2080.

Pei Wu and Hongbing Xing. 2020. The influence of content, lexical and discourse features on the quality of csl learners' writing. *Language Teaching and Research (yǔyán jiāoxué yǔ yánjiū)*, (2):24–32.

Jifeng Wu, Zhou Wei, and Dawei Lu. 2019. Assessing chinese l2 writing quality on basis of language features and content quality. *Chinese Teaching in the World (shìjiè hànyǔ jiāoxué)*, 33(2):130–144.

Jifeng Wu. 2018. The research of indices of the grammatical complexity in south korean native speakers' chinese writing and its relationship with writing quality. *Linguistic Sciences(yǔyán kēxué)*, (5):66–75.

H. Yannakoudakis and R. Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, page 213–223.

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.

Jinming Zhao. 2018. Pedagogical grammar of chinese as a second language:combination of grammar framework and fragmentation. *Language Teaching and Research (yǔyán jiāoxué yǔ yánjiū)*, (2):1–10.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

Appendix

A The Linguistic Complexity and Writing Error Features

We include the detailed list of 90 linguistic complexity and 5 writing error features in Table A1. As described in the main paper, the feature selection module yields two feature sets, i.e. `ling` (33 indices, denoted as ◇) and `ling+err` (31 indices, denoted as ♣).

Table A1: The feature sets in this study.

ID	Feature	Description
Chinese characters and vocabulary		
1	CHAR_NUM ◇ ♣	number of Chinese characters
2	WORD_NUM ◇	number of words
3	LEXICAL_RTTR ◇ ♣	Root type token ratio (RTTR) of words
4	LEXICAL_SOP2 ◇ ♣	Root ratio of sophisticated words
Sentences and clauses		
5	MLS	Mean length of sentences
6	MLC ◇ ♣	Mean length of clauses
7	MLTU	Mean length of T-units
8	NCPS	Number of clauses per sentence
9	NTPS	Number of T-units per sentence
10	MEAN_TREE_DEPTH	Mean depth of syntactic trees
11	MAX_TREE_DEPTH ◇ ♣	Max depth of syntactic trees
Collocations and bigrams		
12	COLL_RTTR	RTTR of all the collocations
13	UNIQUE_RTTR	RTTR of Chinese unique collocations
14	GENERAL_RTTR ◇	RTTR of language-independent collocations
15	UNIQUE_RATIO2 ◇	Ratio of Chinese unique collocations
16	LOWFREQ_RATIO2 ◇ ♣	Ratio of sophisticated collocations
17	VO_RATIO ♣	Ratio of verb-object collocations
18	VO_RTTR ◇	RTTR of verb-object collocations
19	SP_RATIO ♣	Ratio of subject-predicate collocations
20	SP_RTTR ◇	RTTR of subject-predicate collocations
21	AN_RATIO	Ratio of adjective-noun collocations
22	AN_RTTR	RTTR of adjective-noun collocations
23	AP_RATIO ◇	Ratio of adverb-predicate collocations
24	AP_RTTR ♣	RTTR of adverb-predicate collocations
25	CN*_RATIO ◇ ♣	Ratio of classifier-noun collocations
26	CN*_RTTR ◇ ♣	RTTR of classifier-noun collocations
27	PP*_RATIO ♣	Ratio of preposition-postposition collocations
28	PP*_RTTR ◇	RTTR of preposition-postposition collocations
29	PV*_RATIO	Ratio of preposition-verb collocations
30	PV*_RTTR ◇ ♣	RTTR of preposition-verb collocations
31	PC*_RATIO	Ratio of predicate-complement collocations
32	PC*_RTTR	RTTR of predicate-complement collocations
33	BIGRAM_RTTR ◇	RTTR of bigrams
34	BIGRAM_SOP2 ◇	Root ratio of sophisticated bigrams
Dependency structures		
35	DEP_RTTR	RTTR of dependency triples
36	DEP_SOP2 ◇ ♣	Root ratio of sophisticated dependency triples

Continued on next page

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

Table A1 – continued from previous page

ID	Feature	Description
37	HED_RTTR	RTTR of HED dependency triples
38	HED_RATIO	Ratio of HED dependency triples
39	COO_RTTR ◇ ♣	RTTR of COO dependency triples
40	COO_RATIO ◇ ♣	Ratio of COO dependency triples
41	SBV_RTTR	RTTR of SBV dependency triples
42	SBV_RATIO ◇	Ratio of SBV dependency triples
43	ADV_RTTR	RTTR of ADV dependency triples
44	ADV_RATIO	Ratio of ADV dependency triples
45	ATT_RTTR ◇	RTTR of ATT dependency triples
46	ATT_RATIO ♣	Ratio of ATT dependency triples
47	VOB_RTTR	RTTR of VOB dependency triples
48	VOB_RATIO ◇ ♣	Ratio of VOB dependency triples
49	FOB_RTTR	RTTR of FOB dependency triples
50	FOB_RATIO	Ratio of FOB dependency triples
51	POB_RTTR	RTTR of POB dependency triples
52	POB_RATIO ◇ ♣	Ratio of POB dependency triples
53	IOB_RTTR ♣	RTTR of IOB dependency triples
54	IOB_RATIO	Ratio of IOB dependency triples
55	DBL_RTTR ♣	RTTR of DBL dependency triples
56	DBL_RATIO ◇	Ratio of DBL dependency triples
57	RAD_RTTR	RTTR of RAD dependency triples
58	RAD_RATIO	Ratio of RAD dependency triples
59	CMP_RTTR	RTTR of CMP dependency triples
60	CMP_RATIO	Ratio of CMP dependency triples
61	LAD_RTTR	RTTR of LAD dependency triples
62	LAD_RATIO ♣	Ratio of LAD dependency triples
63	COO_DIST	Mean distance of COO dependency triples
64	SBV_DIST	Mean distance of SBV dependency triples
65	ADV_DIST	Mean distance of ADV dependency triples
66	ATT_DIST ◇ ♣	Mean distance of ATT dependency triples
67	VOB_DIST ◇ ♣	Mean distance of VOB dependency triples
68	FOB_DIST	Mean distance of FOB dependency triples
69	POB_DIST ◇	Mean distance of POB dependency triples
70	IOB_DIST ◇	Mean distance of IOB dependency triples
71	DBL_DIST	Mean distance of DBL dependency triples
72	RAD_DIST	Mean distance of RAD dependency triples
73	CMP_DIST	Mean distance of CMP dependency triples
74	LAD_DIST	Mean distance of LAD dependency triples
75	MEAN_DIST	Mean distance of all the dependency triples
Constructions		
76	CONST_DENSITY ◇ ♣	Number of constructions / number of characters
77	CONST1_RATIO ◇ ♣	Ratio of level-1 constructions
78	CONST1_DENSITY ◇ ♣	Number of level-1 constructions / number of characters
79	CONST2_RATIO	Ratio of level-2 constructions
80	CONST2_DENSITY	Number of level-2 constructions / number of characters
81	CONST3_RATIO	Ratio of level-3 constructions
82	CONST3_DENSITY	Number of level-3 constructions / number of characters
83	CONST4_RATIO	Ratio of level-4 constructions

Continued on next page

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

Table A1 – continued from previous page

ID	Feature	Description
84	CONST4_DENSITY	Number of level-4 constructions / number of characters
85	CONST5_RATIO	Ratio of level-5 constructions
86	CONST5_DENSITY	Number of level-5 constructions / number of characters
87	CONST_LOW_RATIO	RATIO of low level constructions (level-1 and level-2)
88	CONST_HIGH_RATIO	RATIO of high level constructions (level-4 and level-5)
89	CONST_LOW_DENSITY	Number of low level constructions / number of characters
90	CONST_HIGH_DENSITY	Number of high level constructions / number of characters
Writing errors		
91	PUNC_ERROR_NUM ♣	Number of punctuation errors
92	CHAR_ERROR_NUM ♣	Number of character level errors
93	WORD_ERROR_NUM ♣	Number of word level errors
94	SENT_ERROR_NUM ♣	Number of sentence level errors
95	DISCOURSE_ERROR_NUM	Number of discourse level errors

B The Example Essays

Essay 1 (high, 95 points)

Translation of Prompt: How should we view euthanasia?

- 作文题：如何看待“安乐死”

在二十世纪的今天，“安乐死”已不是什么新话题。在不同的国家，已曾有人要求、助他人安乐死；只是每个国家的法律有，因此对助他人求安乐死的人的对待亦有所不同。比方说，根据香港的法律，自和人均属违法。法院不判犯人死刑；自的人如果自不遂，救後活过来，理论上是要坐牢的。

法律归法律，竟生与死是人一生的大事，不能绝对由法律。代人崇尚自由，什么都讲求选择——生育女、职业、居住地、个人格……偏偏在出生件事情上，没有一个人可以选择。那么，人最少应该可以选择何时死亡了吧?

讨论“安乐死”的基础，应该是人对生命的尊重。要不然，有了安乐死为後盾，人可以在痛苦中言死去，而存心作奸犯科的，更大有理由害命。

我为“安乐死”有其可取之处，但得以可为大前提。首先，只有身患重病，到了末期段，亦明知没有医治方法的病人，有权选择“安乐死”。第二，这个选择必须由病人在清醒的下自作出，最好有书面证明。第三，关於病情的判断，应有最少位医生签署证明作实。有些人或还会加上第四个条件，就是只准许以「被动」方式施行“安乐死”，即拔掉维生器具；不能施行「主动」方式的“安乐死”，即注射毒药等。

“安乐死”可以免除一些受痛苦煎熬的病人的困苦，使他们仍有一点尊严地去世，有其可取之处；但防止有人滥用“安乐死”，亦致为重要。

Essay 2 (medium, 65 points)

Translation of Prompt: How to solve the so-called generation gap problem?

- 作文题：如何解决“代沟”问题

从古代到在，代沟问题是在人们的生活上常存的。着社会的发展速度加快，代沟问题的程度也很深了。那么我们如何解决这个问题呢?

对我的意见来说，最好的解决方法就是增加两代之间的对话。为了增加对话呢，需要两代互相的努力。首先找个共同的话题开始慢慢得增加在一起的时间。

举例子说，我常用电脑的。无论工作还是玩儿，都来电脑做的。但是我妈对我这个样子太不满意了。这也是一种代沟吧。妈妈是对电脑外行，但孩子每天用电脑所以她对我不懂的地方也越来越多。她还不满意吗？我妈跟我经过一段时间的对话，我才了解了她的心情是如何。然后

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1202-1217, Hohhot, China, Augest 13 - 15, 2021.

我开始教妈电脑。在呢我妈跟我一起用电脑。如果我对妈妈的态度只是不满意没有跟她说话的话，我解决不了这个问题。我感觉得这样的办法对孩子也有教育方面的好处。

总而言之，我想代沟的问题呢，应该通过两代互相的努力增加对话时间的时候，可以解决的。

Essay 3 (high, 45 points)

Translation of Prompt: How do I view popular songs?

- 作文题：我看流行歌曲

我非常喜欢流行歌曲，因为流行歌曲不但动听，而且可以表达自己的想法和感情。比如说：周杰伦、王力宏、田震、那英、周传雄等等他们都是发自自己内心再唱歌。

还有那些作词、作曲的人都用音乐来表达自己或其他人的情感，对社会赞赏和不满。

流行歌曲里大部分都情歌，一个人想对自己喜欢的人告白，用歌曲是最好不过的了。

希望人人都喜欢流行歌曲。