Overview of CCL23-Eval Task 8: Chinese Essay Fluency Evaluation (CEFE) Task

Xinshu Shen[1], Hongyi Wu[1], Man Lan[1,2,*], Xiaopeng Bai[2,3], Yuanbin Wu[1,2], Aimin Zhou[1,2], Shaoguang Mao[4], Tao Ge[4] and **Yan Xia[4]**
[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Institute of AI for Education, East China Normal University
[3]Department of Chinese Language and Literature, East China Normal University
[4]Microsoft Research Asia
{xinshushen,hongyiwu}@stu.ecnu.edu.cn
{mlan,ybwu,amzhou}@cs.ecnu.edu.cn,xpbai@zhwx.ecnu.edu.cn
{shaoguang.mao,tage,yanxia}@microsoft.com

Abstract

This paper provides a comprehensive review of the CCL23-Eval Task 8, i.e., ***C**hinese **E**ssay **F**luency **E**valuation* (**CEFE**). The primary aim of this task is to systematically identify the types of grammatical fine-grained errors that affect the readability and coherence of essays written by Chinese primary and secondary school students, and then to suggest suitable corrections to enhance the fluidity of their written expression. This task consists of three distinct tracks: (1) *Coarse-grained and fine-grained error identification*; (2) *Character-level error identification and correction*; (3) *Error sentence rewriting*. In the end, we received 44 completed registration forms, leading to a total of 130 submissions from 11 dedicated participating teams. We present the results of all participants and our analysis of these results. Both the dataset and evaluation tool used in this task are available[1].

1 Introduction

As a life-long and continuous process, education continually evolves and adapts, especially with the widespread of the Internet. Consequently, the task of student essay assessment has considerably broadened in scale. This significant growth has brought the issues of cost-effectiveness and efficiency in manual essay correction to the forefront, marking them as noteworthy considerations in modern educational practices. In response to this, numerous researchers and institutions have begun exploring the potential of computer technology for automated essay correction (Rudner et al., 2006). This initiative serves a dual purpose. First, by analyzing various aspects of an essay, including its language, content, structure, and the challenges that exist within, it allows for the provision of objective, precise, and timely feedback and may equip students with an enriched understanding of their writing challenges, enhancing their overall skills. Second, this allows teachers to effectively gauge students' writing proficiency and provide more targeted guidance, thereby advancing the educational development of students.

Among the key aspects considered during essay correction by teachers is the fluency of expression. The fluency of an essay mirrors the coherence and grammatical correctness of the text, in addition to giving an insight into the author's writing proficiency and expressive capabilities. Enhancing this aspect carries significant implications for improving the quality of essay corrections and elevating the writing standards of the authors themselves.

However, existing evaluation of essay fluency at primary and secondary levels has the following issues: **1) Lack of specifications:** Current work mainly evaluates essay quality overall, with little in-depth research in fluency and a lack of systematic evaluation specifications, which is not beneficial for comprehensive understanding and improvement of students' writing skills. **2) Poor interpretability:** Prior research typically treats fluency as a scoring task (Mim et al., 2021), providing only an overall rating or score. Alternatively, it is treated as a simple grammatical error correction (GEC) task (Gong et al.,

*Corresponding author.
[1]https://github.com/cubenlp/2023CCL_CEFE

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 282-292, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2021; Tsai et al., 2020). These studies focus primarily on identifying and rectifying simple grammatical errors in sentences, examining them through the lens of revisions such as additions, deletions, and modifications. However, these approaches often neglect to study the specific types of grammatical errors and do not indicate the particular error type. However, both detailed error types and correction references are helpful to students, enabling them to understand their mistakes, revise their essays, and avoid the same errors in the future. **3) Lack of data from authentic writing contexts of primary and secondary school students:** Public datasets for researching essay fluency among Chinese primary and secondary school students are scarce, and previous GEC-based research often relies on rule-based or inter-language datasets from Chinese language learners. However, the types of errors found in Chinese students' compositions are more diverse and involve more complex grammar knowledge. Figure 1 provides exemplars of sentences extracted from compositions penned by primary and secondary school students, highlighting their respective errors alongside appropriate corrective suggestions. Usually, an individual sentence encompasses multiple categories of errors that go beyond the confines of mere spelling errors.

Chinese Sentence	English Translation
Sentence：我一共种了两株在阳台上，我平时见不到它们，只有在周末才能望上几眼。 ErrorType：语序不当、主语多余 RevisedSentence：我在阳台上一共种了两株，平时见不到它们，只有在周末才能望上几眼。	Sentence：I planted two plants in total on the balcony. I can't see them usually, only catch a glimpse of them on weekends. ErrorType：Inappropriate Word Order,Subject Redundancy RevisedSentence： I on the balcony planted two plants in total, andcan't see them usually, only catch a glimpse of them on weekends.

Figure 1: An example of our task. In modern Chinese, adverbials are typically positioned between the subject and the predicate rather than at the end of the sentence, thereby leading to an 'Inappropriate Word Order' error. Moreover, in the first two short sentences, there is a problem of 'Subject Redundancy' where the subject 'I' is repeated unnecessarily.

These gaps in existing methodologies underscore the necessity for a fine-grained, interpretable approach that not only identifies errors but also provides detailed, actionable feedback for students, and emphasize the importance of using composition data from authentic writing contexts of primary and secondary school students. Motivated by this, we present the CCL23-Eval task: ***C**hinese **E**ssay **F**luency **E**valuation* (**CEFE**), which aims to identify and correct errors that affect the fluency of writing for primary and secondary school students. The task featured three tracks: (1) *Coarse-grained and fine-grained error identification*; (2) *Character-level error identification and correction*; (3) *Error sentence rewriting*, aiming at providing a higher-quality evaluation of fluency in primary and secondary school essays.

This task attracted 44 teams to sign up for the competition, and in the end, we received 130 submissions from 11 teams. The task description is presented in Section 2. We describe the data we used in this task in Section 3. In Section 4, we discuss the metrics used to rank participant submissions. We list participants' information and results from their submissions and provide a more in-depth discussion in Section 6. In Section 7, we introduce the methods of the excellent teams. We finally conclude the paper in Section 8.

2 Task Description

The mission of our evaluation is categorized into three distinct tracks, each designed to address a specific aspect of identifying and rectifying errors within primary and secondary school compositions. The tasks aim to illuminate the types of errors that students commonly make, thereby providing a foundation for targeted improvements in writing skills.

2.1 Track 1: Coarse-Grained and Fine-Grained Error Identification

Track 1 focuses on the identification of erroneous sentence types in primary and secondary school compositions. Different types of grammatical errors can reflect various writing challenges faced by students, but traditional practices fail to highlight these errors explicitly. This task approaches the issue from two perspectives, character-level and component-level errors, and defines four types of coarse-grained

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 282-292, Harbin, China, August 3 - 5, 2023.

grammatical error types: *Character-Level Error **(CL)***, *Incomplete Component Error **(IC)***, *Redundant Component Error **(RC)***, and *Incorrect Constituent Combination Error **(ICC)***. Furthermore, we have defined fourteen fine-grained error types, which provide a more detailed understanding of the errors that may occur in writing. This task is especially challenging due to the limited writing skills of primary and middle school students, resulting in multiple errors within the same sentence. Therefore, this track is characterized as a multi-label classification task. The detailed descriptions and examples of each types are available on the competition homepage[1], and the detailed category definitions are as follows:

Character-Level Error (CL) Including four fine-grained error types: Word Missing (**WM**), where a word in a commonly used fixed collocation is missing from the sentence and needs to be added; Typographical Error (**TE**), where there are typos in the sentence that need to be revised or deleted; Missing Punctuation (**MP**), where punctuation is missing from the sentence and needs to be added; and Wrong Punctuation (**WP**), where the punctuation used in the sentence is wrong and needs to be revised or deleted.

Redundant Component Error (RC) Three fine-grained error types are: Subject Redundancy (**SR**), which occurs when a complex adverb is followed by a repeated subject referring to the same entity, and the modification is to delete one subject; Particle Redundancy (**PR**) refers to the redundant use of particles, which should be deleted during editing; Other Redundancy (**OR**) refers to any redundant elements not covered by the previous types, which should also be deleted in modification.

Incomplete Component Error (IC) Four fine-grained error types with incomplete components are: Unknown Subject (**US**), which occurs when the sentence lacks a subject or the subject is unclear, and the solution is to add or clarify the subject; Predicate Missing (**PM**) refers to a sentence lacking verbs, which may be corrected by adding predicates; Object Missing (**OBM**) means that a sentence lacks an object, and the solution is to add an object; Other Missing (**OTM**) refers to other missing components besides the incomplete subject, predicate, and object, which may be corrected by adding the missing components except for the subject, predicate, and object.

Incorrect Constituent Combination Error (ICC) Including three fine-grained error types: Inappropriate Verb-Object Collocation (**IVOC**) refers to the predicate and object not being properly matched, and may be corrected by replacing either the predicate or object with other words; Inappropriate Word Order (**IWO**) means that the order of words or clauses in the sentence is unreasonable, and may be corrected by rearranging some words or clauses; Inappropriate Other Collocation (**IOC**) refers to any element in the sentence not covered by the previous types being improperly matched, and may be corrected by replacing it with other words.

2.2 Track 2: Character-Level Error Identification and Correction

Track 2 centers around the recognition and correction of character-level errors in primary and secondary school compositions. These errors primarily fall into four categories: *Word Missing **(WM)***, *Typographical Error **(TE)***, *Missing Punctuation **(MP)***, and *Wrong Punctuation **(WP)***. This track necessitates a composition sentence as input and generates an output in the form of a triplet consisting of the error category and the correction method, including the position of the original error, the operation to be performed (addition-A, replacement-R, or deletion-D) and the result of the modification. Given the multiplicity of error categories, this task also stands as a multi-label classification task.

2.3 Track 3: Error Sentence Rewriting

Track 3 entails the rewriting of incorrect sentences in primary and secondary school compositions. The challenge here is to provide a minimal modification plan for the erroneous sentences while preserving the original semantics. The revision should make as few alterations as possible, as excessive modifications cannot assist students in identifying their writing problems. This concept of automatic sentence correction is vital for teachers to comprehend their students' writing challenges and to consequently improve the students' writing proficiency. It reflects the importance of preserving the student's original thought process while guiding them towards grammatical correctness and clarity of expression.

[1] https://github.com/cubenlp/2023CCL_CEFE

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 282-292, Harbin, China, August 3 - 5, 2023.

	Train Set	Dev Set	Test A	Test B
Track 1	104	27	1,237	4,116
Track 2	103	22	998	4,001
Track 3	100	19	1,236	4,116

Table 1: The statistics of the **CEFE 1.0** corpus and the number of sentences in each of three tracks.

3 Datasets

In a bid to promote and advance research on essay fluency in primary and secondary school students, we annotated both fine-grained grammatical error types and corresponding correction references that affect sentence fluency. We constructed a fine-grained dataset for **C**hinese **E**ssay **F**luency **E**valuation (**CEFE 1.0**), which aims to provide meaningful insights into the nature and types of grammatical errors students typically make in their writing.

3.1 Data Collection

The seed material for the dataset originated from actual compositions written by primary and secondary school students for their exams. The collected data covered various genres of writing, such as character and scene description. This data source was chosen due to its inherent authenticity and richness, emanating from real-world writing scenarios. These exam essays provide authentic and unadulterated insights into the writing abilities, patterns, and common mistakes of students within these age groups. As a result, we were able to encounter a diverse and complex array of error types and revisions, offering a genuine reflection of the challenges students face when writing essays. By grounding our research in these authentic compositions, we ensured that our findings and solutions would remain relevant and applicable to actual student writing, thereby significantly enhancing the potential impact of our work.

3.2 Data Annotation

Annotators, consisting of four undergraduates, four postgraduates majoring in language-related fields, and four expert reviewers with experience as Chinese teachers, were tasked with annotating error types and providing sentence revisions based on error types. The annotation followed the principle of minimal changes. Before actual annotation, annotators received training on the specifications. During the annotation process, the initial annotation was carried out by an undergraduate and a postgraduate student. Following this, expert reviewers conducted a verification pass and made any necessary corrections to ensure the accuracy and reliability of the annotated data. We divided the data into five groups for annotation and held weekly online discussions to address common issues and make adjustments. This dual focus on identifying specific errors and providing correction suggestions not only enhances the interpretability of the task but also empowers students with the necessary understanding to rectify their writing.

3.3 Data Statistics

This section presents the released training and test data for each track. Due to the scarcity of annotated data in real-world scenarios, we require participants to establish high-quality sentence fluency evaluation models on a given small number of samples. The evaluation is divided into two stages, Test A and Test B. The test set may contain correct sentences, and a subset of blind test data is selected for evaluation. We provide standard answers to half of the test data for participating teams to review their own results and conduct in-depth research. The two-phased evaluation design was aimed at optimizing the participating teams' error detection and correction strategies, promoting innovation, and enhancing the overall quality of outcomes in this challenging task. The size of the dataset for each task is shown in Table 1.

4 Evaluation Metrics

We use different evaluation metrics in different tracks of the task. Our precision and recall calculations are the same in all tracks. Precision is defined as the ratio of correctly identified instances to the total

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 282-292, Harbin, China, August 3 - 5, 2023.

number of identified instances. Recall is defined as the ratio of correctly identified instances to the total number of instances labeled in the ground truth. The F1-score, often used in tasks involving binary or multi-class classification, is the harmonic mean of precision and recall, calculated using the formula: $F_1 = \frac{2PR}{P+R}$.

4.1 Track1: Coarse-Grained and Fine-Grained Error Identification

The total score of Track1 is composed of two parts: coarse-grained and fine-grained wrong sentence identification score. The specific calculation method is as follows:

$$F_1 = 0.5 * F_1^{coarse\_grained} + 0.5 * F_1^{fine\_grained} \tag{1}$$

Specifically, precision (P), recall (R), and micro F_1 are used to evaluate the recognition effect of coarse and fine-grained wrong sentence types.

4.2 Track2: Character-Level Error Identification and Correction

The total score for Track 2 is composed of two parts: the score for character-level error type recognition and the score for character-level error correction. The specific calculation method is detailed below (note that correct sentences are not included in these calculations):

$$F_1^{final} = 0.5 * F_1^{identify} + 0.5 * F_1^{correct} \tag{2}$$

4.2.1 Character-Level Error Type Identification Score

We use precision (P), recall (R), micro F_1 to evaluate the recognition effect of character-level error types.

4.2.2 Character-Level Error Correction Score

We also use precision (P), recall (R), and micro F_1 to evaluate the results. Evaluate from word granularity and sentence granularity, the specific calculation method is as follows:

$$F_1^{correct} = 0.8 * F_1^{character\_level} + 0.2 * F_1^{sentence\_level} \tag{3}$$

Each particle size evaluates the result from two parts of detection and correction. The specific calculation method is as follows:

$$F_1^{character\_level} = 0.8 * F_1^{character\_level}(Detection) + 0.2 * F_1^{character\_level}(Correction) \tag{4}$$

$$F_1^{sentence\_level} = 0.8 * F_1^{sentence\_level}(Detection) + 0.2 * F_1^{sentence\_level}(Correction) \tag{5}$$

4.3 Track3: Error Sentence Rewriting

Due to the diversity of rewriting results provided by participants,we evaluate the results of the model from two perspectives, and the top 5 teams in the final rankings will be subject to manual evaluation(correct sentences will not be included in the evaluation):

Comparison with golds We employ three evaluation metrics: **1)** Exact Match (EM): calculates the percentage of correct sentences generated by the model that exactly match the correct references; **2)** Edit metrics proposed by MuCGEC (Zhang et al., 2022): converts error-correct sentence pairs into operations, and compares the model's output operations with the correct references, and calculates the highest scores for precision, recall, and $F_{0.5}$; **3)** BLEU (Papineni et al., 2002): measures the overlap between the model-generated sentences and the correct references.

Correctness and reasonableness of results We also use three evaluation metrics: **1)** Perplexity (PPL): measures the quality of rewritten sentences by BERT (Devlin et al., 2018); **2)** Levenshtein Distance: calculates the edit distance between the rewritten sentence and the original sentence. In composition correction, we aim to transform incorrect sentences into correct ones with as few modifications as possible, as excessive revisions may hinder students' understanding of their mistakes; **3)** BERTScore (Zhang et al., 2019): measures the similarity between the rewritten sentence and the original sentence.

We finally weighted multiple metrics to obtain the final score:

$$FinalScore = (EM + BLEU + F_{0.5} + BERTScore)/4 - Levenshtein - BERT_{PPL} \tag{6}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 282-292, Harbin, China, August 3 - 5, 2023.

ID	Team Name	Organization	Track 1	Track 2	Track 3
1	HIT-SCIR	Harbin Institute of Technology	✓	✓	✓
2	ZUT	Zhongyuan University of Technology	✓	✓	✓
3	ihuman	ihuman	✓	✗	✗
4	HDZ	Individual	✓	✗	✗
5	SEU-SC	Southeast University	✗	✗	✓
6	HIT_2	Harbin Institute of Technology	✗	✗	✓
7	HYY	Individual	✗	✗	✓
8	BLCU-LCC-Lab	Beijing Language and Culture University	✓	✗	✗
9	QT	Individual	✗	✗	✓
10	MBZ	Individual	✗	✗	✓
11	BK	Individual	✗	✗	✓
Total Number		44	36	28	30

Table 2: The basic information of the participants with a total of 44 teams, where 36 teams for Track 1, 28 teams for Track 2 and 30 teams for Track 3.

Team Name	Rank	Final Score	Test	Avg F_1	Coarse-Grained F_1	Fine-Grained F_1
HIT-SCIR	1	52.16	A	47.09	60.18	34.00
			B	52.16	56.70	47.62
ZUT	2	51.96	A	45.89	58.16	33.63
			B	51.96	59.60	44.31
ihuman	3	51.60	A	47.99	61.26	34.71
			B	51.60	58.30	44.89
HDZ	4	49.40	A	-	-	-
			B	49.40	59.99	38.81
Baseline	-	49.40	A	-	-	-
			B	49.40	54.39	44.41
BLCU-LCC-Lab	-	-	A	40.54	53.70	27.38
			B	-	-	-

Table 3: Results of Track 1 *Coarse-Grained and Fine-Grained Error Identification*, where "-" indicates that the team did not submit results. Our baseline model was trained on the training dataset and Test A dataset.

Team Name	Rank	Final Score	Test	Avg F_1	Identify	Correct	Character		Sentence	
							Detection	Correction	Detection	Correction
HIT-SCIR	1	67.33	A	19.99	36.77	3.21	4.00	2.06	1.85	0.58
			B	67.33	74.22	60.44	62.28	62.76	55.86	40.02
ZUT	2	59.85	A	54.42	56.73	52.12	53.49	54.93	48.29	34.32
			B	59.85	67.08	52.61	53.86	55.01	49.81	34.28
Baseline	-	57.81	A	-	-	-	-	-	-	-
			B	57.81	68.76	46.85	47.07	53.33	43.89	29.26

Table 4: Results of Track 2 *Character-Level Error Identification and Correction*. We trained our baseline model using the training dataset and Test A dataset.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 282-292, Harbin, China, August 3 - 5, 2023.

Team Name	Rank	Final Score	Test	$F_{0.5}$	EM	BLEU-4	BERT$_{PPL}$	Levenshtein	BERTScore
HIT-SCIR	1	57.83	A	17.97	7.44	85.85	3.16	3.35	96.57
			B	45.81	17.34	89.85	2.91	1.91	97.60
ZUT	2	56.27	A	42.91	19.42	91.45	3.23	1.24	97.78
			B	40.32	13.03	90.75	2.94	1.23	97.64
SEU-SC	3	55.87	A	42.69	19.26	91.35	3.27	1.19	97.78
			B	39.14	12.58	90.57	2.95	1.16	97.63
Baseline	-	53.40	A	-	-	-	-	-	-
			B	35.32	9.43	89.10	2.96	1.45	97.40
HIT_2	4	53.39	A	33.81	15.13	89.70	3.47	1.75	97.48
			B	34.52	10.87	89.22	2.98	1.65	97.48
HYY	5	52.70	A	17.06	5.18	88.44	3.40	1.00	96.96
			B	30.84	8.45	89.94	3.01	0.97	97.51
MBZ	-	-	A	28.48	7.77	90.43	3.33	0.69	97.73
			B	-	-	-	-	-	-
BK	-	-	A	15.09	4.45	88.45	3.41	0.91	96.90
			B	-	-	-	-	-	-
QT	-	-	A	14.26	4.69	76.74	4.54	7.26	94.63
			B	-	-	-	-	-	-

Table 5: Results for Track 3 *Error Sentence Rewriting*, where "-" indicates that the team did not submit results. Our baseline model was trained using training dataset and Test A dataset.

5 Baselines

We present the outcomes of our baseline models for reference. The training dataset and Test A dataset are utilized for training, and we evaluate the performance of our model on Test B dataset. For Track 1, we fine-tune BERT (Devlin et al., 2018) on our dataset over 100 epochs, employing batch sizes within the range of $[16, 24]$, a learning rate of $2e^{-5}$, and the Adam (Kingma and Ba, 2015) optimizer. Sentences are encoded with these Pretrained Language Models (PLMs) to derive contextual representations (utilizing $[CLS]$ embedding), and error types are identified via fully-connected layers. For Track 3, we fine-tune Chinese BART (Lewis et al., 2019) on our dataset over 100 epochs, with a batch size of 16, a learning rate of $2e^{-5}$, and the AdamW (Loshchilov and Hutter, 2019) optimizer. For Track 2, we employ the model framework of Track 1 and Track 3 to train on Track 2 data, and the error correction result is transformed into the Track 2 format via a script. The results of our baseline models are detailed in Section 6.

6 Results and Analysis

6.1 Results

In our competition, a total of 11 teams submitted their final results. The basic information about them are detailed in Table 2. Ultimately, the performance of the teams was evaluated based on the results from the Test B. It's important to note that any team that did not submit results for this set was not included in the final rankings. The final results for the each track are given in Table 3, Table 4 and Table 5.

6.2 Futher Analysis

Given that our evaluation represents a few-shot task, data augmentation emerges as a prevalent strategy. Moreover, considering the impressive language understanding and generation capabilities of contemporary Large Language Models (LLMs), these models present an effective solution to address such few-shot challenges. Therefore, we counted the use of these two technologies by the participating teams, as shown in Table 6. Based on their results, their performance is superior to the baseline system trained using more

Team Name	used LLMs	used Data Augmentation
HIT-SCIR	✗	✓
ZUT	✗	✗
ihuman	✓	✗
SEU-SC	✗	✓

Table 6: A summary of the methods used by participating teams that contributed implementations. "LLMs" indicates whether the participating team uses large language models; "Data Augmentation" indicates whether to use data outside the task for training.

data, demonstrating the effectiveness of their data augmentation methods and indicating that both data augmentation and using LLMs may effectively increase data quantity, improve model generalization ability, and accuracy.

Moreover, we collected performance metrics for each team on fine-grained error categories and conducted further analysis. Specifically, we assessed the teams' proficiency in identifying 4 coarse-grained and 14 fine-grained error categories, as shown in Figure 2. It may be observed that the models developed by participating teams generally performed well in identifying character-level error categories. However, for more complex grammar error categories such as ICC, the performance is generally less satisfactory whether using rule-based methods or existing CGEC dataset for data augmentation. In other words, there exists a significant discrepancy between the complex grammatical errors present in data constructed using rule-based methods and the actual mistakes made by students in real-world writing scenarios. This further highlights the challenges of our task and underscores the necessity of researching more complex grammar errors that arise in real-world writing scenarios.

For the error sentence rewriting task, the performance of participating teams compared to the standard answers was not ideal, which was shown in Table 5. However, based on metrics such as BERTScore and PPL, the generated sentences were semantically consistent and fluent, according to human cognition of natural language sentences. Existing generation models produce diverse results, but our task aims to correct error sentences on the basis of minimal changes, and this strongly constrained generation requires further exploration.

7 Participant Systems

The participating teams in the task adopted diverse approaches for error detection and correction in primary and secondary school students' essays. This section gives an overview of the methods that have been successful in each of the tracks. Each team's unique approach illustrates the diversity of methods that can be utilized in automated essay assessment and presents various potential directions for future research.

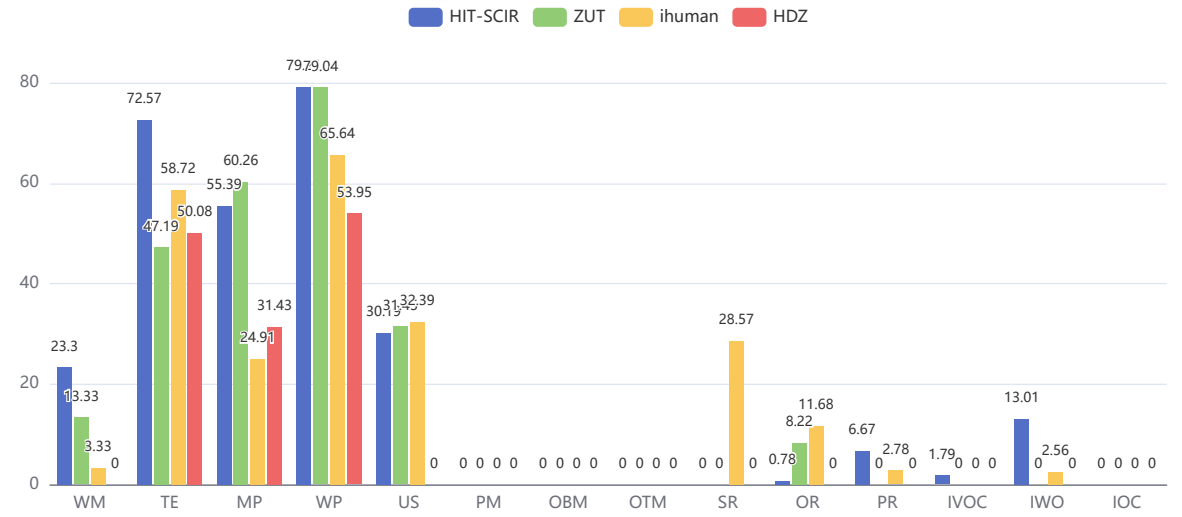
7.1 Track1: Coarse-grained and Fine-grained Error Identification

For Track 1, ***HIT-SCIR*** adopted a fine-grained error detection model based on sequence labeling. Due to the misalignment between the sequence labeling task and the provided human-annotated data, they constructed a large amount of pseudo-data for various types of errors based on LTP (Che et al., 2020) and heuristic rules, which were used for the training of the Track 1 model. At the same time, they used techniques such as model inheritance and threshold post-processing to alleviate the bias caused by pseudo-data training.

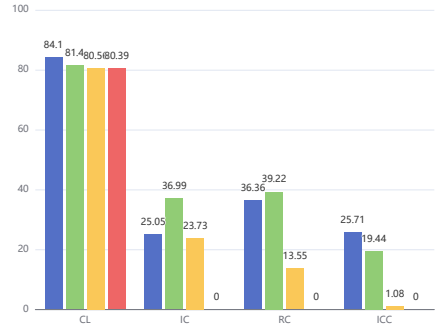
ZUT used the unified heterogeneous supervised multi-task pre-training learning model UTC as the framework. During the fine-tuning process, they incorporated prompt learning, which transformed the multi-classification task into a form similar to cloze-style completion, in order to fully leverage the potential of the pre-trained model.

ihuman directly fine-tuned the ChatGLM-6B (Du et al., 2022) large language pre-training model through LoRa (Hu et al., 2022) technology, and directly used the probability distribution of the output

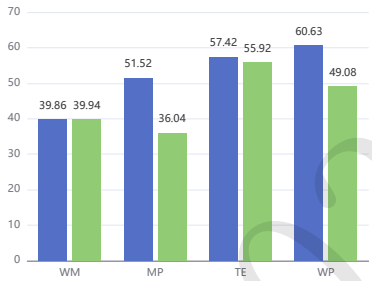
Proceedings of the 22nd China National Conference on Computational Linguistics, pages 282-292, Harbin, China, August 3 - 5, 2023.



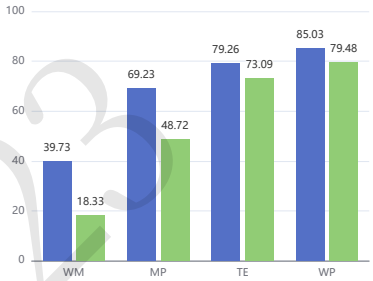
(a) Fine-grained identification results in Track 1.



(b) Coarse-grained results in Track 1.



(c) Identification results in Track 2.



(d) Correction results in Track 2.

Figure 2: (a) shows the fine-grained identification results in Track 1. (b) displays coarse-grained identification results in Track 1. (c) shows the character-level error identification results in Track 2. (d) indicates the character-level error correction results in Track 2. The blue color in the figures represents the results of Team ***HIT-SCIR***, the green color represents the results of Team ***ZUT***, the yellow color represents the results of Team ***ihuman***, and the red color represents the results of Team ***HDZ***.

specific token to predict the result. Key innovative features of their methodology include: foregoing the addition of an extra classification model to enhance the efficiency of model learning; improving the utility of the input information for the same input sentence by concatenating a variety of different prompts; and targeting multiple tokens for model output, thereby enhancing model stability, as well as effectively mitigating the risk of model overfitting.

7.2 Track2: Character-level Error Identification and Correction

For Track2, ***HIT-SCIR*** used the same method in Track 1 for the error identification task and trained an auto-encoder model based on edit label for the character-level error correction based on the pseudo-data mentioned in Track 1. In particular, they constructed more fine-grained pseudo data for character-level errors. They used the GECToR framework (Omelianchuk et al., 2020) based on editing tag sequence labeling to implement addition, deletion, and modification operations for Chinese characters and punctuation errors. ***ZUT*** employed frameworks used on Track 1 and Track 3 to make predictions on the test dataset.

7.3 Track 3: Error Sentence Rewriting

For Track 3, ***HIT-SCIR*** considered using a Seq2Seq model BART (Lewis et al., 2019) to cover the correction with a larger edit distance. They still used the two-stage method of pre-training with pseudo data and fine-tuning with real data to train the error correction model. They used the conventional way

to train the Seq2Seq model and adopted greedy-search decoding in the inference stage to avoid over correction.

ZUT proposed a sequence diffusion process that leverages pre-trained models. By treating the erroneous and correct text as sequences, they designed a classifier-free sequence diffusion process that established connections between two different feature spaces. Additionally, they combined the pre-trained model ERNIE (Sun et al., 2021) with the diffusion model to align decoding ability of ERNIE with the denoising process of the diffusion model, thus achieving text correction capability. As for the dataset, they increased the number of samples by combining data from other tracks in this task to address the issue of insufficient training data.

SEU-SC proposed a model framework consisting of four key modules to address the problems of popular research that often overlooks the utilization of syntactic information and suffers from excessive correction: the data augmentation module, the semantic encoding module, the syntactic encoding module, and the fused information decoding module. To augment the existing Chinese text corpus, a data augmentation approach grounded in syntactic rules and error distribution was employed. This approach strives to amass supplementary training data and enhance the efficacy, generalization, and resilience of the Chinese text correction model. Moreover, the model integrates a graph convolutional network (GCN) (Kipf and Welling, 2016) within the encoder to encode syntax information. The encoded outcomes from the GCN-based syntax information encoder are combined with the encoded outputs from the BART Encoder-based text information encoder. Subsequently, the combined results are fed into the BART Decoder-based decoder to generate grammatically accurate sentences.

8 Conclusions and Future Work

This paper presents an overview of the CCL23-Eval Task ***C**hinese **E**ssay **F**luency **E**valuation* (**CEFE**). We conduct this evaluation using our meticulously annotated **CEFE 1.0** dataset. The evaluation is divided into three distinct tracks: (1) *Coarse-grained and fine-grained error identification*; (2) *Character-level error identification and correction*; (3) *Error sentence rewriting*. Each one aims at addressing a specific facet of grammatical error identification and correction within primary and secondary school compositions. We received a total of 44 completed registration forms, culminating in 130 submissions from 11 participating teams. In addition, we provide a comprehensive analysis and summary of the methodologies employed by the participants, which will contribute to future research in this field of natural language processing. In the future, we will continue to explore methods to improve the identification of fine-grained error types and moderate correction, as well as further investigate the effectiveness of LLMs in our task.

Acknowledgements

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and Human Machine Collaborated Decision Making Methodology (72192820 & 72192824), Pudong New Area Science Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004) and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2020. N-ltp: An open-source neural language technology platform for chinese. *arXiv preprint arXiv:2009.11616*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu. 2021. Iflyea: A chinese essay assessment system with automated rating, review generation, and recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. 2021. Corruption is not all bad: Incorporating discourse structure into pre-training via corruption for essay scoring. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2202–2215.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: tag, not rewrite. *arXiv preprint arXiv:2005.12592*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Lawrence M Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S Chang. 2020. Lingglewrite: a coaching system for essay writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–133.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. Mucgec: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. *arXiv preprint arXiv:2204.10994*.