CCTC: A Cross-Sentence Chinese Text Correction Dataset for Native Speakers

Baoxin Wang[1,2], Xingyi Duan[2], Dayong Wu[2,3], Wanxiang Che[1], Zhigang Chen[2,4], Guoping Hu[2]

[1]Research Center for SCIR, Harbin Institute of Technology, Harbin, China

[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China

[3]iFLYTEK AI Research (Hebei), Langfang, China

[4]Jilin Kexun Information Technology Co., Ltd., Changchun, China

{bxwang2,xyduan,dywu2,zgchen,gphu}@iflytek.com

car@ir.hit.edu.cn

Abstract

The Chinese text correction (CTC) focuses on detecting and correcting Chinese spelling errors and grammatical errors. Most existing datasets of Chinese spelling check (CSC) and Chinese grammatical error correction (GEC) are focused on a single sentence written by Chinese-as-a-second-language (CSL) learners. We find that errors caused by native speakers differ significantly from those produced by non-native speakers. These differences make it inappropriate to use the existing test sets directly to evaluate text correction systems for native speakers. Some errors also require the cross-sentence information to be identified and corrected. In this paper, we propose a cross-sentence Chinese text correction dataset for native speakers. Concretely, we manually annotated 1,500 texts written by native speakers. The dataset consists of 30,811 sentences and more than 1,000,000 Chinese characters. It contains four types of errors: spelling errors, redundant words, missing words, and word ordering errors. We also test some state-of-the-art models on the dataset. The experimental results show that even the model with the best performance is 20 points lower than humans, which indicates that there is still much room for improvement. We hope that the new dataset can fill the gap in cross-sentence text correction for native Chinese speakers.

non-native examples

WRONG: 父母对孩子的爱情在世界上是最重要的。
CORRECT: 父母对孩子的关爱在世界上是最重要的。
TRANSLATION: The **love** of parents for their children is the most important in the world.

WRONG: 这一点可以说是吸烟对个人健康的利益。
CORRECT: 这一点可以说是吸烟对个人健康的好处。
TRANSLATION: This can be said to be a **benefit** of smoking to the health of the individual.

native examples

WRONG: 弹奏者只有做到手臂、肘部与腕部都能够完全随着指尖动作运行，才能有信息弹奏出均匀、流畅的音符，音色也会更加出挑。
CORRECT: 弹奏者只有做到手臂、肘部与腕部都能够完全随着指尖动作运行，才能有信心弹奏出均匀、流畅的音符，音色也会更加出挑。
TRANSLATION: Only when the player is able to run his arms, elbows and wrists completely with the fingertip movements can he have the ~~information~~ **confidence** to play even, smooth notes and have a more distinguished tone.

Figure 1: Comparison between the errors caused by native and non-native speakers. The non-native examples are from CGED 2018, and the native examples are from CCTC.

1 Introduction

Chinese text correction (CTC) aims at detecting and correcting errors in Chinese text. Text correction has important applications in the domain of education, journalism, and publishing. For many native Chinese speakers, such as journalists, writers, and bloggers, a text correction system for native Chinese speakers will greatly improve the efficiency of their proofreading. In the field of NLP, Chinese text corrections usually includes two tasks: Chinese spelling check (CSC) (Hong et al., 2019; Cheng et al., 2020; Wang et al., 2021) and Chinese grammatical error correction (GEC) (Yuan and Briscoe, 2016; Omelianchuk et al., 2020; Wang et al., 2020).

The existing CSC and Chinese GEC test sets (Tseng et al., 2015; Rao et al., 2018; Zhao et al., 2018) are mainly generated from essays written by Chinese-as-a-second-language (CSL) learners. The essays written by CSL learners are significantly different from those written by native Chinese speakers. Specifically, essays written by CSL learners usually contain more errors and are more likely to make mistakes in the misuse of words. In contrast, texts produced by native speakers contain sparser errors and typically make mistakes that are caused by oversight. These significant differences prevent researchers from using the existing test sets directly to evaluate text correction systems for native speakers.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3331–3341
October 12–17, 2022.

Figure 1 shows the errors made by CSL learners and native speakers, respectively. We can see the CSL learners make some mistakes that are obvious to native speakers. The word “爱情” usually refers to the love between a couple, while “关爱” indicates the love of an elder for a younger child. In Chinese, these two words are not interchangeable. However, For CSL learners, it is easy to mistakenly write “关爱” as “爱情” because they can both be translated into “love” in English. Similarly, the words “利益” and “好处” can both be translated into “benefit” in English, but the word “利益” cannot be used with “健康” (health) in Chinese. Native speakers will not make these mistakes. For native Chinese speakers, the most common errors are caused by oversight, which the writers themselves are capable of correcting. For example, the misspelling of “信息”(information) as “信心” (confidence) is due to the similarity of the Pinyin for *xinxi* and *xinxin*, respectively. Besides, the test sets for non-native speakers, such as CGED (Rao et al., 2018), and SIGHAN (Tseng et al., 2015) tend to write simpler sentences with limited topics. In contrast, the texts written by native speakers tend to have complicated sentences with various topics.

Moreover, the existing datasets of CSC and GEC are mainly for sentence-level correction. However, some errors usually need to be corrected via the cross-sentence information (Chollampatt et al., 2019; Yuan and Bryant, 2021). For example, in Figure 2, it is difficult to see what is wrong with each sentence individually. According to the previous sentences, we know that the word “蜘蛛” (spider) should be corrected as “红蜘蛛” (red spider).

WRONG:	**红蜘蛛**俗称火蜘蛛、火龙。红蜘蛛**……**。危害特点：蜘蛛是一种危害作物种类较多的害虫，以成虫、幼虫或若虫群聚在叶背吸取汁液。
CORRECT:	**红蜘蛛**俗称火蜘蛛、火龙。红蜘蛛**……**。危害特点：红蜘蛛是一种危害作物种类较多的害虫，以成虫、幼虫或若虫群聚在叶背吸取汁液。
TRANSLATION:	**Red spider** is commonly known as fire spider and fire dragon. Red spider … . Damage characteristics: ~~Spider~~ **Red spider** is a pest that affects more crop species, with adults, larvae, or worm clusters in the back of the leaves to suck sap.

Figure 2: An example for cross-sentence text correction.

To better evaluate the text correction system’s performance on document-level texts produced by native speakers, we propose a new dataset CCTC (**C**ross-Sentence **C**hinese **T**ext **C**orrection). Since every Chinese character may be erroneous, the scale of annotation is large. Without any auxiliary hints, the annotators will be prone to miss the errors. Therefore, we give the annotators some hints about the position and type of errors produced by several CSC and GEC systems. We first annotate all the sentences from 200 documents and find only 11.4% sentences with errors. Errors in sentences with candidate errors account for more than 90% of all errors. In order to maximize the diversity of topics and increase the number of errors in the dataset, we only annotate the sentences with error candidates for another 1,300 documents. Concretely, we annotate 1,500 texts from the Internet, and the annotated text includes a total of 30,811 sentences and more than 1 million Chinese characters.

We utilize several types of state-of-the-art models for experiments and analyses on our dataset. We also evaluate the performance of native speakers on CCTC. The experimental results show that even the model with the best performance is still 20 points worse than the human, which indicates that there is still much room for improvement.

To summarize, our contributions are as follows:

- We propose a new Chinese text correction dataset, which can be used to evaluate text correction systems for native speakers better.
- Our dataset is based on document-level text. We have done some experiments and analyses for cross-sentence errors, which we hope will be helpful for subsequent studies of cross-sentence text correction.
- We systematically compare our dataset with other CSC and GEC datasets and test four state-of-the-art models on the new dataset.

We hope that CCTC will contribute towards the development of cross-sentence Chinese text correction for native speakers. Our datasets are publicly available at `https://github.com/destwang/CTCResources`.

2 Existing Datasets

The Chinese text correction related datasets mainly include Chinese spelling check (CSC) and grammatical error correction (GEC). Statistics information is shown in Table 1, and the features of these datasets are shown in Appendix.

2.1 English GEC Datasets

CoNLL14 The test set (Ng et al., 2014) consists of essays written by English as a Second Language

Datasets	# sents	Avg. Sent. Length	Avg. Doc. Length	Err. Sent. (%)	Sent-K	# tokens	Language	Task
CoNLL 2014	1,312	22.9	-	75.8	0.25	30,045	En	GEC
JFLEG	747	18.9	-	86.4	0.53	14,118	En	GEC
CWEB-S	2,864	23.9	-	24.5	0.39	68,450	En	GEC
CWEB-G	3,981	20.3	-	25.6	0.44	80,814	En	GEC
SIGHAN 2015	1,100	30.5	-	50.0	-	33,550	Zh	CSC
OCR Text	1,000	10.2	-	100.0	-	10,198	Zh	CSC
CGED 2018	3,549	39.6	-	56.0	-	140,655	Zh	GEC
NLPCC 2018 GEC	2,000	29	-	99.2	-	59,325	Zh	GEC
CCTC-Train	12,689	41.9	818.6	9.8	0.76	532,088	Zh	CTC
CCTC-W	14,338	38.8	856.6	9.4	0.72	556,767	Zh	CTC
CCTC-H	3,784	41.4	784.2	11.4	0.78	156,836	Zh	CTC

Table 1: Statistics of datasets. For datasets CGED 2018, NLPCC 2018 GEC and SIGHAN 2015, the statistics here are about their test sets. All test sets are sentence-level except for our dataset CCTC. Here, tokens mean the subwords obtained after tokenizing of BERT, which are mainly individual Chinese characters for Chinese. Sent-K is Cohen's Kappa at sentence level. CCTC-H means a high-quality test set, and CCTC-W means a test dataset which contains a wider range of documents.

(ESL) learners from the National University of Singapore, which are annotated for grammatical errors by two native English speakers.

JFLEG The JFLEG corpus (Napoles et al., 2017) consists of sentences written by English language learners for the TOEFL exam. The texts have been corrected for grammatical errors and fluency.

CWEB This dataset (Flachs et al., 2020) is designed to annotate English web text, which corresponds to a dataset containing both native and non-native speakers.

CWEB is the closest to our proposed dataset among the known datasets. There are three main differences: (i) our dataset is document-level, while CWEB is sentence-level; (ii) our data only focus on the texts written by native speakers; (iii) our proposed dataset is designed for Chinese.

2.2 CSC Datasets

SIGHAN 2015 The text of SIGHAN 2015 (Tseng et al., 2015) is collected from the essay section of the computer-based Test of Chinese as a Foreign Language (TOCFL). Thus, the spelling errors are mainly caused by CSL Learners. SIGHAN 2015 is based on the sentence, and the rate of the erroneous sentences is manually adjusted to be higher than the original text.

OCR Text The dataset is produced from OCR results of Chinese subtitles in videos (Hong et al., 2019). Therefore, these sentences are from native Chinese speakers, but these errors are automatically generated by the OCR method and not caused by human writing.

2.3 Chinese GEC Datasets

CGED 2018 The corpora used in CGED 2018 (Rao et al., 2018) are taken from the writing section of the HSK (*Hanyu Shuiping Kaoshi*, Pinyin of "A test of Chinese level"). The grammatical errors are also produced by non-native speakers. There are four kinds of errors, which are spelling errors, redundant words, missing words, and word ordering errors.

NLPCC 2018 GEC The training data (Zhao et al., 2018) is mainly collected from Lang-8. The test data is extracted from the PKU Chinese Learner Corpus, which is constructed by the Department of Chinese Language and Literature, Peking University.

MuCGEC The dataset consists of 7,063 sentences collected from CSL learner sources. MuCGEC (Zhang et al., 2022) is a multi-reference multi-source evaluation dataset for Chinese Grammatical Error Correction.

In contrast to CGED 2018, NLPCC 2018 GEC and MuCGEC datasets, CCTC is based on document-level texts written by native speakers.

3 CCTC Dataset

We construct a new cross-sentence Chinese text correction dataset for native speakers. We extract the raw text from WuDaoCorpora (Yuan et al., 2021), which mainly includes news, blogs, and some popular science articles. We pre-process the collected documents, remove personal information, advertisements, and noisy articles, then sample 1,500 documents for annotation. We take 100 of these documents for verification. We can determine by the author's information that all the 100 documents

Candidate Methods	# sents	# err. sents
BERT-CSC	3,905	2,192
BERT-GEC	2,213	1,404
BERT-CGED	2,083	356
Others	2,734	40
Total	10,935	3,992

Table 2: Statistics of different candidate generation methods. *# sents* is the number of candidates generated by these methods and *# err. sents* is the number of real erroneous sentences. *Others* indicates the sentences which are labeled without error candidates in CCTC-H.

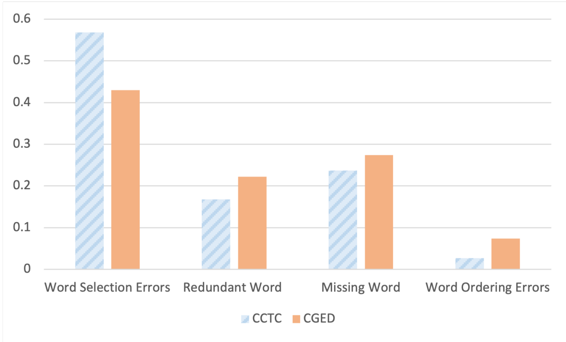

Figure 3: The rate of different error types.

are written by native Chinese speakers, which illustrates that almost all of the documents are written by native Chinese speakers. Table 1 shows the statistical information.

Candidates Generation To facilitate manual annotation and reduce error omission, we utilize several different models to generate error candidates. Specifically, we select three different kinds of models as follows. The detailed information of the training set will be described in the next section.

- BERT-CSC: We train a BERT-based (Devlin et al., 2019) Chinese spelling check model via the pseudo-data similar to Cheng et al. (2020).
- BERT-GEC: We replace, insert, delete and shuffle some tokens randomly to construct GEC pseudo-data and train a BERT-based sequence labeling model.
- BERT-CGED: We train a BERT-based sequence labeling model using the CGED training dataset.

To cover as many errors as possible, we lower the thresholds of the three models. In this way, these models will generate more candidates to find out the erroneous parts of the documents.

Annotation Following Rao et al. (2018), errors are divided into four types: spelling errors (word selection errors), redundant words, missing words, and word ordering errors. The data are annotated by five annotators, with an average of about 120 hours and 2K sentences each. Our annotators annotate the dataset on an annotation tool prepared in advance. We pay our annotators appropriately according to the number of annotated sentences.

We firstly annotate 3,784 sentences from 200 documents, including sentences with error candidates and sentences without candidates. After annotating, we find that there are only 431 sentences with errors. Errors in candidate sentences account for more than 90% of all errors. In order to maximize the diversity of topics and increase the number of errors in the dataset, we only annotate the sentences with error candidates for another 1,300 documents. We name the dataset with 200 annotated documents as **CCTC-H**, which means a high-quality dataset. The remaining 1,300 documents are divided into two parts, 650 of which are used as the training set and the other 650 documents as the **CCTC-W**, which means this test dataset contains a wider range of documents. To conclude, we annotate 1,500 documents from the Internet, and the annotated texts include a total of 30,811 sentences and more than 1 million Chinese characters. The detailed statistics of different candidate generation methods are shown in Table 2.

In order to ensure the quality of the annotated data, we take 500 sentences from the training set, validation set, and test set, respectively, and annotate these sentences without candidate errors. Similar to Flachs et al. (2020), annotator agreement is calculated at the sentence level using Cohen's Kappa. Kappa is 0.76, 0.72, and 0.78 for the CCTC-Train, CCTC-W, and CCTC-H, respectively, showing that our dataset has a higher agreement than the previous dataset.

Dataset Analysis Table 3 shows examples of the four types of errors. Figure 3 shows the rate of sentences corresponding to the four error types. We can see that Chinese spelling errors (word selection errors) are the most common in documents written by native speakers, accounting for about 60% of the total. Word ordering errors have the least percentage of all errors. For texts written by non-native speakers from CGED, redundant words and missing words occur at a relatively greater rate than texts written by native speakers. The occurrence of

Error Type	Example sentence	Translation
Spelling Errors	进入大学，就是进入一个新的环境，~~结出~~（接触）新的人，你的所有过去对于他们来说是一张白纸。	Entering college means entering a new environment, you will ~~bear~~ (meet) new people, and all your past is a blank sheet of paper to them.
Redundant Words	突然有一天，一个女人来看~~来看~~孩子。	Suddenly one day, a woman came to see ~~came to see~~ the child.
Missing Words	今天要讲（的）是他在一年时间里面的教师生涯。	What today is going (to) be talking about is his career as a teacher inside a year.
Word Ordering Errors	一般室内环境采用200系列材质即可，而室外~~需环境~~（环境需）使用304材质。	General indoor environment needs to use 200 series material, while outdoor ~~needs environment~~ (environment needs) to use 304 series material.

Table 3: Examples of different error types caused by native speakers.

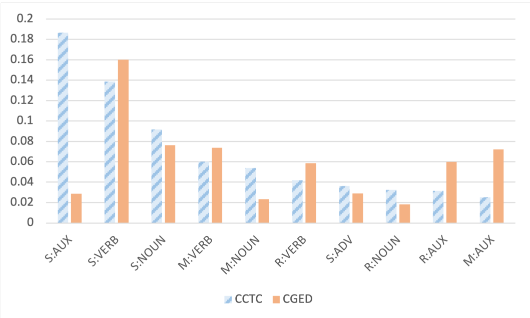

Figure 4: The rate of different error types with POS tagging. We count the multi-token errors according to the POS tags of multi-token. If the multi-token errors can be segmented into k words, the count of each type will increase by $1/k$. (S: Spelling errors, M: missing words, R: redundant words)

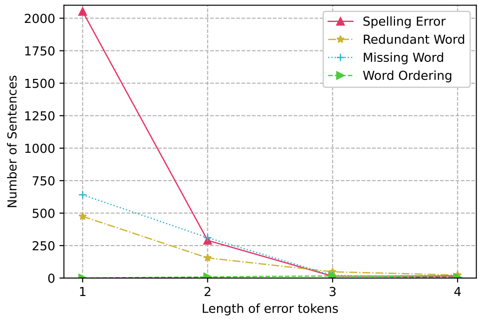

Figure 5: The length of error span.

word ordering errors is rare for native speakers and somewhat more frequent for non-native speakers.

To better analyze the difference between errors made by native and non-native Chinese speakers, we perform statistical lexical analysis for each error type. In this paper, we use LTP (Che et al., 2010) for the Part-of-Speech (POS) tagging of the text. The statistical results are shown in Figure 4. We find that the most common mistake made by native speakers is the misuse of auxiliaries. In contrast, non-native speakers tend to write a sentence with redundant or missing auxiliaries.

We count the length of the error span, which can be seen in Figure 5. Except for the word ordering errors, the errors with one token are in the majority. The decline in the percentage of spelling errors of two consecutive tokens is faster than the percentage for redundant and missing words. Errors of more than three consecutive tokens are rare.

We perform a manual statistical analysis of the dataset and find that 68% of errors are caused by oversight, such as spelling errors caused by the Pinyin Input method. The word "接触" (meet) may be incorrectly entered as "结出" (bear) due to similar pronunciation as shown in Table 3. This type of error is varied, making this type of error more difficult to correct. The remaining errors are mainly due to misuse of some words with similar semantics or method of use, such as the auxiliaries "的" and "地". In Chinese, "的" is usually used as a suffix of adjective and "地" is used as a suffix of adverb, and they are pronounced the same, so these two words are often misused in Chinese.

We analyze the spelling errors more specifically. The spelling errors can be divided into the following five types: misuse of words, single Chinese character error in a word, pronoun errors, auxiliary errors, and other single Chinese character errors, accounting for 28%, 23%, 8%, 30%, and 11%, respectively.

4 Experiments

4.1 Training Dataset

There are no training datasets specifically annotated for errors caused by native speakers before.

Dataset	# sents	err. sents (%)
CGED	44,754	94.7
NLPCC GEC	1,200,000	89.8
SIGHAN	281,381	100.0
Pseudo-data	3,000,000	99.6

Table 4: Statistics of training dataset.

In this paper, in addition to our proposed training set, we use training data from multiple sources, including CGED (Rao et al., 2018), SIGHAN (Tseng et al., 2015), and NLPCC 2018 GEC dataset (Zhao et al., 2018). For the CGED data[1], we use CGED training data from 2014 to 2016, totaling about 45K sentences. For NLPCC dataset[2], there are multiple correction sentences for each sentence. We randomly select part of the correction sentences as our training set. For SIGHAN, we use the training data of SIGHAN, as well as the automatically generated corpus (Wang et al., 2018). Besides, we also use our training set of CCTC to train these models. For the GECToR model, we only use CCTC to fine-tune after the pseudo-data training the same as Omelianchuk et al. (2020).

As mentioned above, native speakers make a wider variety of errors, so we use heuristics to construct pseudo-data in the hope that we can cover as many types of errors as possible. We construct a large-scale pseudo-data using Chinese Wikipedia. The pseudo-data generation method for GEC is similar to Zhao et al. (2019), which randomly delete, add, replace, and shuffle the tokens. To better check the Chinese spelling errors, for the replacement operation, 80% of the tokens are from the confusion set provided by Wu et al. (2013) and 20% of the tokens are from the corpus. The pseudo-data of CSC are generated by the same replacement operation. Table 4 shows the statistics of the training data.

4.2 Models

We evaluate performance on our proposed dataset using four state-of-the-art approaches to CSC or GEC. The specific models are described as follows.

- SpellGCN (Cheng et al., 2020): This model incorporates phonological and visual similarity knowledge into BERT via a specialized graph convolutional network.
- ResBERT (Wang et al., 2020): ResBERT is the state-of-the-art model in CGED competition, by adding ResNet to the BERT model to achieve better performance.
- GECToR (Omelianchuk et al., 2020): GECToR achieves the correction of errors such as redundant words, missing words, and spelling errors by the BERT model.
- CopyNet (Zhao et al., 2019): CopyNet is a transformer-based seq2seq model, which can pay more attention to the grammatical errors through the copy mechanism.

4.3 Metrics

In the previous works, GEC systems are usually evaluated using F0.5-score based on MaxMatch (Dahlmeier and Ng, 2012) since that the precision of the GEC system is more critical for ESL or CSL learners. On the contrary, recall is usually more important than precision for native Chinese speakers because most errors are caused due to oversights. They can make correct judgments about most grammatical errors by themselves. Therefore, we use the F2-score to evaluate the performance on the CCTC dataset. The specific equation is as follows:

$$\text{F2-score} = \frac{5 \times \text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}} \quad (1)$$

Given that native speakers can generally make correct judgments by themselves, it is also essential for them to detect the position of errors as well. Regarding CGED (Rao et al., 2018), SIGHAN (Tseng et al., 2015), and NLPCC (Zhao et al., 2018), we perform three kinds of evaluation, namely sentence-level, position-level, and correction-level evaluation. The sentence-level evaluation determines whether there is an error in a sentence, while the position-level evaluation needs to label the error position correctly. For the correction-level evaluation, we statistically score the systems by the error position, error type, and correction results similar to Rao et al. (2018). The difference is that we use F2-score because the recall for native speakers is usually more important.

4.4 Experimental Settings

We use the RoBERTa-wwm (Cui et al., 2019) as the base models of SpellGCN, GECToR, and ResBERT. The training hyperparameters of SpellGCN and CopyNet are kept consistent with Cheng et al.

[1] http://www.cged.tech
[2] http://tcci.ccf.org.cn/conference/2018/taskdata.php

Test Set	Train Set	Model	Sentence-Level P	Sentence-Level R	Sentence-Level F2	Position-Level P	Position-Level R	Position-Level F2	Correction-level P	Correction-level R	Correction-level F2
CCTC-W	SIGHAN	SpellGCN	23.59	41.66	36.12	10.99	21.88	18.26	9.49	18.89	15.77
	CGED	ResBERT	15.96	73.14	42.61	6.50	33.50	18.30	-	-	-
	NLPCC	GECToR	27.90	30.90	30.25	8.38	10.57	10.04	7.29	9.20	8.74
	Pseudo-data	CopyNet	14.04	**78.75**	40.98	1.59	16.22	5.72	0.90	9.14	3.22
		ResBERT	26.13	40.61	36.56	11.34	20.01	17.36	-	-	-
		GECToR	26.29	44.39	39.02	11.61	22.25	18.80	8.17	15.66	13.24
	CCTC-Train	SpellGCN	**55.61**	43.48	**45.46**	**38.96**	31.44	**32.71**	**35.19**	**28.40**	**29.54**
		ResBERT	17.62	49.89	36.51	13.38	**37.65**	27.63	-	-	-
		GECToR	43.13	45.88	45.30	23.37	26.26	25.63	20.36	22.87	22.32
CCTC-H	SIGHAN	SpellGCN	26.27	36.71	34.01	11.51	18.10	16.24	10.81	17.00	15.26
	CGED	ResBERT	19.24	64.44	43.83	7.59	**29.07**	18.56	-	-	-
	NLPCC	GECToR	32.55	29.25	29.86	9.58	10.05	9.96	8.07	12.25	11.10
	Pseudo-data	CopyNet	18.55	**79.73**	**48.04**	2.13	17.37	7.13	1.10	8.96	3.68
		ResBERT	26.02	33.08	31.37	9.65	14.26	13.02	-	-	-
		GECToR	27.82	37.67	35.18	10.84	16.45	14.91	8.07	12.25	11.10
	CCTC-Train	SpellGCN	**61.33**	35.80	39.05	**40.44**	23.68	**25.82**	**36.07**	**21.12**	**23.03**
		ResBERT	25.86	39.49	35.72	16.84	25.93	23.40	-	-	-
		GECToR	49.87	33.86	36.19	24.66	17.28	18.38	22.60	15.84	16.85

Table 5: Experimental Result. For the GECToR model, we use CCTC-Train to fine-tune after the pseudo-data training the same as Omelianchuk et al. (2020).

(2020) and Zhao et al. (2019) respectively. For ResBERT, we use the BIO encoding (Kim et al., 2004) the same as Wang et al. (2020). We fine-tune the models using the sentences with errors in CCTC-train.

For CopyNet and GECToR, they will generate a corrected sentence. To evaluate the performance of position-level detection for the two models, we use the Levenshtein[3] distance to convert the sentence pairs into the corresponding error types. Concretely, Levenshtein distance can generate three types of operations: delete, insert and replace, which correspond to redundant words, missing words, and spelling errors. Then we convert the adjacent insertion and deletion operations into word ordering errors. In this way, we can evaluate the detection performance of the two models. Since spelling errors accounted for the highest percentage of all errors, we also test directly using the SpellGCN model, which can only correct the spelling errors.

4.5 Experimental Result

The experimental results are shown in Table 5. The overall performance of the models after training with the CCTC-Train is better than other datasets. CopyNet trained with pseudo-data achieves the best performance for sentence-level detection on CCTC-H. For all the models without CCTC-train, ResBERT with CGED dataset achieves the best results on position-level detection. However, ResBERT only detects the errors, but it cannot correct the sentence. Surprisingly, SpellGCN performs best for correction. This may be because spelling errors account for most errors, and SpellGCN is better able to correct them using phonological and visual similarity knowledge. ResBERT trained with CGED dataset achieves better performance than the model using pseudo-data. We find that ResBERT with CGED is more effective in detecting auxiliary errors such as the misuse of "的" and "地", which account for a relatively large proportion of all errors.

[3]https://github.com/ztane/python-Levenshtein

Model	Sentence-Level P	Sentence-Level R	Sentence-Level F2	Position-Level P	Position-Level R	Position-Level F2
SpellGCN	75.0	33.3	37.5	42.1	20.5	22.9
GECToR	71.4	41.7	45.5	34.8	20.5	22.4
ResBERT	48.0	47.4	47.5	31.7	33.8	33.4
Human	**85.4**	**67.3**	**70.3**	**61.7**	**56.0**	**57.1**

Table 6: Experimental results for comparison with humans. The results of humans are the average results of two untrained native speakers. All the models are trained with CCTC-Train dataset.

Besides, we can see that the precision of each model is higher overall on CCTC-H than on CCTC-W, and the recall is lower. This may be because all sentences in CCTC-H are labeled, and the coverage of errors is greater.

4.6 Analysis

To better evaluate the effectiveness of these models, we test the performance of humans for text correc-

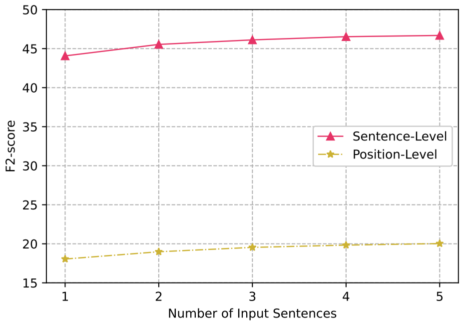

Figure 6: Experimental results for different input sequence length in inference stage, the model is a single-sentence trained ResBERT model.

tion. The low error density in the actual text makes it very difficult for humans to correct texts. Thus, we take 200 sentences from the CCTC-H dataset and adjust the erroneous sentences to about 50%. Two untrained native speakers are asked to correct these 200 sentences. We want to know what performance the native Chinese speaker can achieve. The corresponding experimental results are shown in Table 6. More detailed results are in the Appendix.

After increasing the error density, the performance of almost all the models improves. Human performs much better than these models. Even the model with the best results is 20 points worse than the human, indicating that the models still have much room for improvement.

Also, with the human test, native speakers often miss errors without being informed of the error position in advance, even though we have increased the error rate to about 50%. For example, in Figure 7, an annotator missed the error "有限" (limited) because this word also appears frequently. When we point out this position, native speakers can easily correct the error.

5 Cross-Sentence Errors

We randomly analyze 100 errors and find that cross-sentence information is necessary for only 11% of the errors. However, cross-sentence information can be helpful for 38% of errors, such as when the corrected word appears in context.

To test the help of cross-sentence information for Chinese text correction, we try a simple cross-sentence correction method, which increases the length of the input sequence. We vary from single-sentence correction to multi-sentence correction, and Figure 6 shows the experimental results. From

WRONG: 用户只能使用支持VR视频的三星手机观看，还要下载NBC体育应用并验证自己是NBC有限电视或卫星电视的订户。

CORRECT: 用户只能使用支持VR视频的三星手机观看，还要下载NBC体育应用并验证自己是NBC有线电视或卫星电视的订户。

TRANSLATION: Users can only use VR video-enabled Samsung phones to watch, and need to download the NBC Sports app and verify that they are subscribers to NBC ~~limited~~ **cable** or satellite TV.

WRONG: 天生桥位于今香格里拉县城东10公里处，...... 。作品中人与自然和谐而存的主题也是天生与自然的特点。

CORRECT: 天生桥位于今香格里拉县城东10公里处，...... 。作品中人与自然和谐而存的主题也是天生桥自然的特点。

TRANSLATION: **Tiansheng Bridge** is located 10 kilometers east of Shangri-La County, ... The theme of harmony between man and nature in the work is also a natural feature of ~~natural~~ **Tiansheng Bridge**.

Figure 7: Examples of CCTC. The above sentence is an example of failure to correct during human testing, and the below one is an example for mis-correction by SpellGCN.

the experimental results, we can see that for a trained model, the performance of the model increases as the input sequence length grows. This also shows that the cross-sentence information is helpful for Chinese text correction.

The models often mis-correct some low-frequency words due to the lack of context of a document. In Figure 7, the model mistakenly modify "天生桥自然" (Tiansheng Bridge) as "天生与自然" (Natural). In fact, the word "天生桥" has appeared many times in the context of the document. If we could better use the cross-sentence contextual information, it would help better with the correction. Based on this, we do not simply split the document into individual sentences but keep the complete cross-sentence information. We hope it will be helpful for subsequent studies of cross-sentence text correction.

6 Conclusion

In this paper, we propose a novel cross-sentence Chinese text correction dataset for native speakers. Concretely, we manually annotated 1,500 Chinese texts written by native speakers collected from the Internet. The new dataset consists of 30,811 sentences and more than 1,000,000 Chinese characters. It contains spelling errors, redundant words, missing words, and word ordering errors. CSC and GEC systems developed for native speakers can be better evaluated on CCTC than the previous datasets. We also test some state-of-the-art models on the dataset. The experimental results show that even the model with the best performance is still 20 points worse than the human, which indicates that there is still much room for improvement.

Acknowledgements

This work was supported by the National Key R&D Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 61976072 and 62176078. I would like to thank the anonymous reviewers and Xiaoxue Wang for their insightful comments and suggestions.

References

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese language technology platform. In *Coling 2010: Demonstrations*, pages 13–16, Beijing, China. Coling 2010 Organizing Committee.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 871–881, Online. Association for Computational Linguistics.

Shamil Chollampatt, Weiqi Wang, and Hwee Tou Ng. 2019. Cross-sentence grammatical error correction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 435–445, Florence, Italy. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. 2020. Grammatical error correction in low error density domains: A new benchmark and analyses. *arXiv preprint arXiv:2010.07574*.

Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. 2019. FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 160–169, Hong Kong, China. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. *arXiv preprint arXiv:1702.04066*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170. Association for Computational Linguistics.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of nlptea-2018 share task chinese grammatical error diagnosis. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–51.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 32–37, Beijing, China. Association for Computational Linguistics.

Baoxin Wang, Wanxiang Che, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2021. Dynamic connected networks for Chinese spelling check. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2437–2446, Online. Association for Computational Linguistics.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for Chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages

2517–2527, Brussels, Belgium. Association for Computational Linguistics.

Shaolei Wang, Baoxin Wang, Jiefu Gong, Zhongyuan Wang, Xiao Hu, Xingyi Duan, Zizhuo Shen, Gang Yue, Ruiji Fu, Dayong Wu, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2020. Combining ResNet and transformer for Chinese grammatical error diagnosis. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 36–43, Suzhou, China. Association for Computational Linguistics.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at SIGHAN bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 35–42, Nagoya, Japan. Asian Federation of Natural Language Processing.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California. Association for Computational Linguistics.

Zheng Yuan and Christopher Bryant. 2021. Document-level grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 75–84, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States. Association for Computational Linguistics.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the nlpcc 2018 shared task: Grammatical error correction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 439–445. Springer.

Model	Sentence-Level			Position-Level			Correction-Level		
	P	R	F2	P	R	F2	P	R	F2
SpellGCN	75.0	33.3	37.5	42.1	20.5	22.9	39.5	19.2	21.4
GECToR	71.4	41.7	45.5	34.8	20.5	22.4	32.6	19.2	20.9
ResBERT	48.0	47.4	47.5	31.7	33.8	33.4	-	-	-
Human	**85.4**	**67.3**	**70.3**	**61.7**	**56.0**	**57.1**	**52.2**	**46.2**	**47.2**

Table 7: Experimental results for comparison with humans.

Dataset	Native Speakers	Real Errors	Original Distribution	Cross-Sentence	Grammatical Error
CoNLL 2014		✓			✓
JFLEG		✓			✓
CWEB	-	✓	✓		✓
SIGHAN 2015		✓			
OCR Text	✓				
CGED 2018		✓			✓
NLPCC 2018 GEC		✓			✓
CCTC (Ours)	✓	✓	✓	✓	✓

Table 8: The features of different datasets. The CWEB dataset contains sentence produced by both native English speakers and non-native English speakers. In contrast, our dataset CCTC only contains text written by native Chinese speakers.

A Appendix

Table 7 shows the correction-level experimental results for comparison with humans. Table 8 shows the features of different datasets.