A New Dataset and Empirical Study for Sentence Simplification in Chinese

Shiping Yang[*†] **, Renliang Sun**[*]**, Xiaojun Wan**
Wangxuan Institute of Computer Technology, Peking University
Center for Data Science, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
yangshipingnlp@gmail.com
sunrenliang@stu.pku.edu.cn
wanxiaojun@pku.edu.cn

Abstract

Sentence Simplification is a valuable technique that can benefit language learners and children a lot. However, current research focuses more on English sentence simplification. The development of Chinese sentence simplification is relatively slow due to the lack of data. To alleviate this limitation, this paper introduces CSS, a new dataset for assessing sentence simplification in Chinese. We collect manual simplifications from human annotators and perform data analysis to show the difference between English and Chinese sentence simplifications. Furthermore, we test several unsupervised and zero/few-shot learning methods on CSS and analyze the automatic evaluation and human evaluation results. In the end, we explore whether Large Language Models can serve as high-quality Chinese sentence simplification systems by evaluating them on CSS.

1 Introduction

Sentence Simplification (SS) is the task of modifying a sentence to make it easier to understand and improve its accessibility for a wider audience, while retaining most of its original meaning (Alva-Manchego et al., 2020b). Automatic SS systems provide reading assistance to children (De Belder and Moens, 2010; Kajiwara et al., 2013), non-native readers (Paetzold and Specia, 2016b), and people with reading disabilities (Carroll et al., 1998; Rello et al., 2013; Evans et al., 2014).

Currently, there are multiple datasets for English sentence simplification to choose from, such as WikiSmall (Zhu et al., 2010), WikiLarge (Zhang and Lapata, 2017) and Newsela (Xu et al., 2015). However, few simplification-specific datasets are available for other languages. Research on other popular languages like Spanish (Štajner et al., 2015; Saggion et al., 2015), French (Gala et al., 2020), Russian (Sakhovskiy et al., 2021; Dmitrieva et al., 2021) and Italian (Brunato et al., 2015; Tonelli et al., 2016) has gained momentum in recent years. Unfortunately, research on Chinese sentence simplification is still scarce: to the best of our knowledge, there is currently no publicly available simplification corpora for training, even lacking a dataset to evaluate the ability of simplification models. These limitations greatly prevent the development of Chinese sentence simplification.

Creating a parallel dataset for Chinese sentence simplification poses considerable challenges. Most datasets for SS use the automatic sentence alignment method to construct (Zhang and Lapata, 2017; Dmitrieva et al., 2021), which relies on the existence of large-scale simplification corpus like Simple English Wikipedia[1]. However, there are no suitable data sources in Chinese. Another possible option is to translate the English dataset to Chinese through neural machine translation (Sakhovskiy et al., 2021; Li et al., 2022). Nevertheless, Chinese is different from English in both grammatical structure and language habits, which leads to a significant difference in text simplification between these two languages. Simplifying sentences with the help of human experts is also an option, like it was done in Newsela or ALECTOR[2](Gala et al., 2020), but this is expensive and slow. Due to the above reasons, we decided to manually construct a dataset only for evaluation, achieving a trade-off between cost and feasibility.

In this study, we annotate and propose CSS (**C**hinese **S**entence **S**implification dataset), a new Chinese dataset for the evaluation of SS models. We then apply several unsupervised SS methods and zero/few-shot learning methods on the dataset and analyze the advantages and disadvantages of the methods for Chinese sentence simplification. Furthermore, we study the behavior of popular met-

[*] Equal contribution
[†] Work done during internship

[1] https://dumps.wikimedia.org/simplewiki
[2] ALECTOR is a SS dataset for evaluation in French.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8306–8321
July 9-14, 2023 ©2023 Association for Computational Linguistics

rics for English SS when using them to evaluate simplifications produced by Chinese SS systems.

We are committed to promoting research on Chinese sentence simplification. In general, our contributions can be summarized as follows:

- We create a high-quality dataset named CSS for the evaluation of Chinese SS models. We will publicly release CSS at `https://github.com/maybenotime/CSS`, it will be the first open-source simplification-specific dataset in Chinese.

- We conduct data analysis to compare the characteristics of CSS with English datasets, pointing out the difference between Chinese and English sentence simplification tasks.

- We report the performance of several unsupervised methods and zero-shot/few-shot learning methods on our dataset, which could serve as the baselines for future studies.

2 Related Work

2.1 Sentence Simplification

Sentence simplification research has achieved promising progress in recent years. EditNTS (Dong et al., 2019) simplifies a sentence with iterative explicit editing. ACCESS (Martin et al., 2020a) performs controllable simplification by conditioning specific control tokens. Omelianchuk et al. (2021) proposed a simple and efficient simplification system based on sequence Tagging. However, these methods all relied on supervised parallel training corpora.

To overcome the scarcity of parallel SS corpus in low-resource languages, recent research has proposed many unsupervised methods to train simplification models without a labeled simplification corpus (Kajiwara and Komachi, 2018; Surya et al., 2019; Katsuta and Yamamoto, 2019; Aprosio et al., 2019; Kumar et al., 2020). MUSS (Martin et al., 2020b) obtains strong performance in French and Spanish, even outperforming the supervised state of the art. Lu et al. (2021) further improved the performance by building pseudo-SS corpora with an unsupervised approach. Finally, the experiments of multi-task learning (Dmitrieva and Tiedemann, 2021) and cross-lingual learning (Mallinson et al., 2020) in sentence simplification shows the possibility of performing zero- and few-shot simplification without any parallel data, driving us to explore the Chinese SS task in the zero- and few-shot setting.

2.2 Simplification Datasets in Multiple Languages

There exist a lot of supervised training corpora (Xu et al., 2015; Zhang and Lapata, 2017) and high-quality test datasets (Xu et al., 2016; Sulem et al., 2018a; Alva-Manchego et al., 2020a) for English SS. However, automatic SS systems in other popular languages also have extensive demand and application values. Researchers attempted to explore simplification in other languages (Aluísio et al., 2008; Saggion et al., 2015; Kajiwara and Komachi, 2018), but are limited by the lack of parallel corpora.

Recently, some works have focused on building SS datasets in other low-resource languages (Brunato et al., 2015; Battisti et al., 2020; Sakhovskiy et al., 2021; Dmitrieva et al., 2021) to facilitate the development of multilingual SS techniques, such as ALECTOR (Gala et al., 2020), Simpitiki (Tonelli et al., 2016), and Spanish part of Newsela (Xu et al., 2015). However, to our best knowledge, there is no work attempting to build a Chinese SS dataset, which hinders the development of Chinese SS systems.

3 CSS

In this section, we describe detailed information about our CSS dataset. Specifically, we first give the annotation process of the CSS dataset in Section 3.1. Then, we show statistical information about our CSS dataset in Section 3.2. In Sections 3.3 and 3.4, we do automatic and manual data analysis on CSS. And an additional dataset for few-shot setting is described in Section 3.5.

3.1 Data Collection and Annotation

To obtain the raw texts for CSS, we randomly sampled original sentences from the PFR Chinese corpus. Then, we asked the annotators who have passed the qualification test to simplify those sentences. Except for manual simplifications, annotators were also asked to give the rewriting transformations they performed on the original sentences. We introduce the PFR corpus and Preprocessing details in Appendix A.

Operations Defined According to the previous studies on human simplification (Petersen and Ostendorf, 2007; Feng, 2008), we define 4 simplification operations that can be performed in CSS: (1) lexical simplification (replacing complex words with synonyms or explaining idioms with a short sentence). (2) sentence splitting. (3) compression (deleting unimportant information). (4) sentence paraphrasing (word reordering or syntactic transformations).

Worker Requirements The requirements for workers are as follows: (1) native speakers of Chinese; (2) had education experience in university with at least a bachelor's degree; (3) passed the corresponding Qualification Test designed for our task (more details below). These requirements were designed to ensure the workers have a proficient level of Chinese, and are capable of performing the simplification task.

Qualification Test We designed a Qualification Test (QT) to measure the worker's simplification ability. The content of the test consisted of simplification operation recognition, specific sentence simplification, and free simplification. Before the QT, we showed them detailed explanations of the sentence simplification task and examples of multiple simplification operations we defined. After the annotators took the QT, all submissions were manually checked to filter out workers who could not perform the task correctly. We had 10 candidates take the QT, out of which 5 passed the test and entered the following annotation stage.

Annotation Round Workers who passed the QT would have access to this round. In addition to the simplification of each sentence, workers were also asked to annotate the simplification operations they performed on sentences and submit confidence scores (similar to ASSET (Alva-Manchego et al., 2020a)) on their simplifications using a five-point Likert scale (1:No Simplification Implemented, 2:Very Low, 5:Very High). We finally collected two simplified counterparts for each original sentence from different workers to fit the scenario with multiple transformations. Thus, our dataset is suitable for automatic evaluation metrics that require multiple references, such as BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016).

Simplification Guide Before the QT and the Annotation Round, workers were asked to read the task guide about how to do sentence simplification. We provided examples and definitions of lexical simplification, sentence splitting, compression, and sentence paraphrasing. We also included an example where all transformations were performed. To stimulate their creativity and motivation, we informed workers that they can simplify original sentences with which type of simplification operations at their discretion. Additionally, we added bonuses to encourage workers to perform multiple transformations in a sentence.

Quality Control We added some fake examples to ensure the quality of the dataset. The fake examples were assigned to every worker in each annotation round. We checked whether the workers gave reasonable simplifications by comparing their simplification results with the golden reference. Besides, we manually checked the instance (an original-simplified sentence pair) with a confidence score of 1 and removed the original sentences that have no need to simplify from the dataset.

Table 1 presents an example of simplifications in CSS, together with corresponding translation and operation tags. Please refer to Appendix B for more examples.

3.2 Statistics

CSS consists of 766 human simplifications associated with the 383 original sentences from the PFR corpus (two simplifications per original sentence). Table 2 presents basic statistics of CSS. We also show the same statistics of two mainstream English SS datasets for comparison.

Compared with previous English SS datasets, CSS offers fewer references for each original sentence but with rich rewriting transformations and additional information. While HSplit (Sulem et al., 2018a) contains simplifications produced mostly by sentence splitting, simplifications in CSS are more similar to ASSET's which involve multiple types of rewriting transformations. CSS also provides additional simplification operation tags to show which types of rewriting transformations have been performed on this sentence, different from other datasets. Operation tags can provide help in the evaluation of controlled SS systems.

Table 3 shows the percentage of each simplification operation applied to the original sentences. It can be seen that most of the sentences in Chinese can be simplified by lexical simplification, and few annotators tended to simplify a sentence by sentence splitting. Compression and sentence paraphrasing are also common ways for Chinese SS.

Original	五台县位于山西省忻州市，因境内有五台山而得名，而五台山位居全国四大佛教名山之首。 Wutai County, located in Xinzhou City, Shanxi Province, is named after the Wutai Mountains, which ranks first among the four most famous Buddhist mountains in China.
Reference	五台县在山西省忻州市，因为它里面有五台山从而得到这个名字。 五台山在全国四大佛教名山中排名第一。 Wutai County is in Xinzhou City, Shanxi Province, and got this name because of the Wutai Mountain. Wutai Mountain ranks first among the four most famous Buddhist mountains in China.
Operations	Lexical simplification; Sentence splitting; Sentence paraphrasing

Table 1: Simplification example with corresponding translations and operation tags in CSS.

	CSS	ASSET	HSplit
Ori. Sentences	383	359	359
Num. of Ref.	2	10	4
Multi Operations	√	√	×
Operation Tag	√	×	×
Tokens per Ref.	47.29	19.04	25.49

Table 2: Basic statistics of CSS compared with ASSET and HSplit. From here on, we only report the statistics of the test set of ASSET for a fair comparison.

3.3 Dataset Analysis

We further study the simplifications collected for CSS through a series of surface and syntax-based features.

- **Number of simplification operations:** The number of simplification operations on the simplification instance.

- **Number of sentence splits:** The number of sentences in the simplification minus the number of sentences in the original sentence.

- **Compression level:** The number of characters in the simplification divided by the number of characters in the original sentence.

- **Replace-only Levenshtein distance:** We report the Replace-only Levenshtein distance as described in ASSET, which is computed as character-level Levenshtein distance (Levenshtein et al., 1966) only with replace operations divided by the length of the shortest string. Therefore, this feature serves as a proxy for lexical simplification.

- **Proportion of words deleted, added and reordered:** Number of words deleted/reordered[3] from the original sentence divided by the number of words in the original sentence; and the number of words that were added to the original sentence divided by the number of words in the simplification.

- **Word deletion only:** A boolean feature shows whether the simplification is obtained only by deleting words from the original sentence. This feature captures extractive compression.

- **Lexical complexity score ratio:** Word ranks (in a frequency table) have been shown to be the best indicator of word complexity (Paetzold and Specia, 2016a). We compute the log of word ranks as log-ranks, and then obtain the lexical complexity score[4] by computing the mean squared log-ranks of words in a sentence (without stopwords). We use the Chinese common words frequency table, released by BLCU Corpus Center[5]. The ratio is then the lexical complexity score on the simplification divided by that of the original sentence.

- **Dependency tree depth ratio:** We compute the ratio of the depth of the dependency parse tree of the simplification relative to that of the original sentence. When a simplification contains more than one sentence, we use the maximum depth of all dependency trees as the depth of the simplification. This feature is a good indicator to show the simplification in sentence structure.

Figure 1 shows the density histograms of the features of CSS except **Number of sentence splits** and **Word deletion only**. For some key features that significantly demonstrate the difference between Chinese and English SS datasets, we high-

[3] A reordered word is a word that is contained in the original sentence and simplification but not in the longest common subsequence.

[4] There is a difference in the computing way of lexical complexity score between the source code of *tseval* and description in ASSET(Alva-Manchego et al., 2020a), we following the version of the original paper.

[5] http://bcc.blcu.edu.cn/

Operation	Lexical simplification	Sentence splitting	Compression	Sentence paraphrasing
Percentage(%)	91%	20%	45%	60%

Table 3: the percentage of each simplification operation applied to the original sentences, which is calculated by the number of the original sentences that have a certain operation occurring in references divided by the number of the original sentences in CSS. To some extent, this value can indicate which simplification operation is applicable to most sentences.

	CSS	ASSET
Sentence Splitting	11.8%	20.2%
Compression(<75%)	9.1%	31.2%
Word Reordering	17.6%	28.3%
Word Deletion Only	5.6%	4.5%

Table 4: The percentage of sentences that: have at least one sentence split, have a compression level of 75% or lower, have at least one reordered word, and operate word deletion only.

light these statistics as percentages in Table 4, and report the statistics of ASSET as a comparison.

Nearly half of the instances in CSS perform more than one simplification operation, which shows the diversity of simplification operations in CSS. By analyzing the replace-only Levenshtein distance and the proportion of words deleted/added, we can see how much the annotators have paraphrased the original sentence. Both CSS and ASSET have a very low ratio of word deletion only, which means that few extractive compression operations were performed.

Sentence splitting is a common operation in English sentence simplification. Although annotators in ASSET tended to not split sentences, they still performed sentence splitting at a rate of 20.2% (Alva-Manchego et al., 2020a), and this rate can even reach 68.2% in HSplit. However, the percentage of sentence splitting is only 11.8 in CSS. A reasonable explanation for this phenomenon is that complex English sentences usually contain many nested clauses. Annotators may simplify these sentences by splitting the clause. And a complex Chinese sentence is usually constituted by many short sentences instead of nested clauses. This explanation is complemented by the distributions of dependency tree depth and the percentage of reordered words. In summary, **Chinese SS do fewer structural changes than English.**

We introduce compression operation to simplify sentences in CSS, same with ASSET. However, Table 4 shows that the compression ratio on CSS is much lower than on ASSET. CSS has a high density of a compression ratio of 1.0, even has many instances with compression levels greater than 1.0. This phenomenon can be explained by the frequent use of idioms. In Chinese, an idiom often alludes to a story or historical quotation, compressing a lot of information. It is difficult to simplify idioms just by replacing words. Annotators usually used a short sentence to explain an idiom, which leads to the above phenomenon. **The lexical simplification of Chinese is different from English because of the existence of idioms.**

3.4 Human Rating of CSS

In this section, we measure the quality of the CSS dataset using human judges. Workers needed to satisfy the same basic requirements as described in Section 3.1, and passed the Qualification Test that was designed for human evaluation. Following Alva-Manchego et al. (2020a), we rated the quality of simplifications based on three criteria: simplicity, fluency (or grammaticality), and meaning. Simplicity is the most important indicator in this task.

We invited three workers to evaluate the quality of the CSS dataset with the above criteria. We randomly chose 100 original sentences from the dataset and, for each of them, we sampled one manual simplification. Workers were asked to use the five-point Likert scale to submit their level of agreement (1:Strongly disagree, 5:Strongly agree) with the following statements:

- **Simplicity**: The simplified sentence is simpler and easier to understand than the original sentence.
- **Fluency**: The simplified sentence is fluent and free of grammatical errors.
- **Meaning**: The simplified sentence adequately preserves the meaning of the original, perhaps omitting the least important information.

The average simplicity score is 3.88, indicating the simplification of the original sentence is good

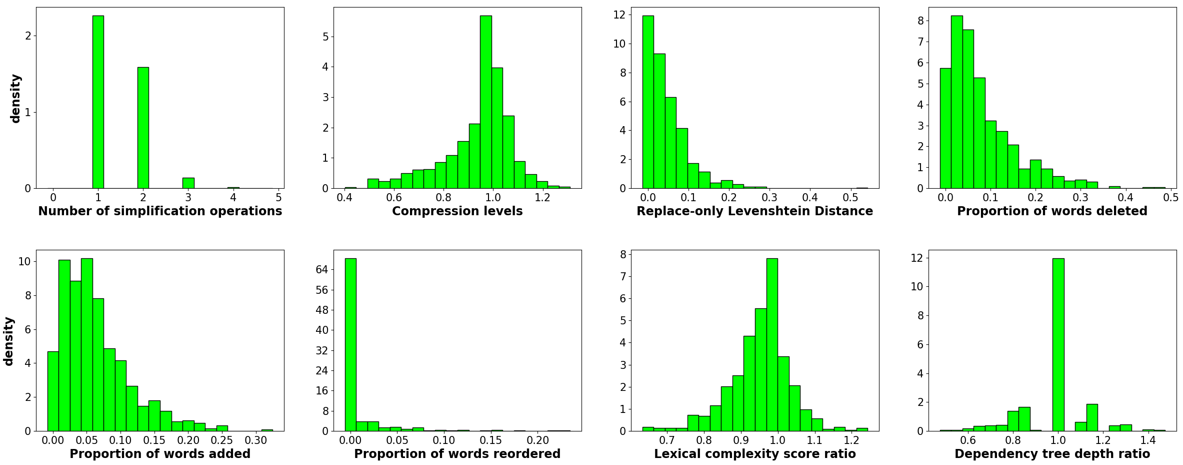

Figure 1: Density of text features in simplifications from CSS.

enough. The average fluency scores reach 4.83, probably because the simplified sentences are written by humans and are easy to read. The meaning score achieves 4.71, implying that the references express the meaning of the original sentences precisely. In all, the quality of CSS is guaranteed.

3.5 Additional Dataset for Few-shot Setting

We annotated an additional dataset following the annotation process described in Section 3.1. Different from CSS, this dataset only consists of 288 manual simplifications, with only one reference for each original sentence. We use part of the dataset as the validation set in our experiment.

The ability to efficiently learn from limited data is critical for NLP tasks. Recently, zero- and few-shot learning with large-scale pre-trained language models have achieved promising progress on generative tasks (Vu et al., 2022). We released this additional dataset to facilitate future works on few-shot sentence simplification. Researchers are free to split the dataset into training and validation sets, using them in the few-shot scenario, and evaluate on CSS.

4 Experiments

In this section, we conducted a series of experiments to explore how to train Chinese SS models in low-resource scenarios. Some of them could serve as the baselines for future studies. All the models are tested on CSS.

Following Martin et al. (2020b), We first implement the following simple methods for comparison.

Identity It simply outputs the original sentence, which means the original sentence and the simplification are exactly the same.

Truncation The original sentence is truncated and only the first 80 percent of words are retained.

Gold Reference We report gold reference scores in CSS as two references are available. We compute scores of one manual simplification, using another simplification as a reference. The scores are then averaged over all references.

We then introduce several unsupervised methods and zero/few-short methods for comparison.

4.1 Unsupervised Method

Automatic SS systems rely on unsupervised techniques when supervised training data is unavailable. Lu et al. (2021) proposed an unsupervised simplification[6] method to build a SS parallel corpus based on a large-scale bilingual translation corpus, which has achieved state-of-the-art results in multiple languages. We replicated this current unsupervised state-of-the-art model in Chinese.

Specifically, we chose English as the bridge language and used News-commentary *en-zh* dataset[7] as the high-resource bilingual translation corpus. Then, we used a machine translation model[8] to translate the English sentences to Chinese. The

[6] Previous works (Martin et al., 2020b) used the term *unsupervised simplification* to describe works that do not use any labeled parallel simplification data While using some supervised components.

[7] News-commentary is a common dataset in the field of neural machine translation, we download it from `https://data.statmt.org/news-commentary/v15/training/`.

[8] `https://huggingface.co/Helsinki-NLP/opus-mt-en-zh`

source sentences (Chinese) and the translated sentences constituted pseudo-complex-simple sentence pairs. Different from the original work (Lu et al., 2021), We filtered pseudo-SS corpus only by BLEU because FKGL metric is not suitable for Chinese.

To compare with the model trained by pseudo-SS data, We provide **translate training** that the original sentence and simplification all are translated from the English WikiLarge dataset as a baseline. We use the same translation model and data size to make a fair comparison.

4.2 Zero- and Few-shot Transfer

In addition to unsupervised methods, recent works on zero- and few-shot learning with pre-trained language models can provide a potential solution for performing the SS task in a low-resource language (Mallinson et al., 2020). We conduct experiments to explore whether the model can obtain prior SS knowledge through **cross-lingual** transfer and **cross-task** transfer. And all the models are trained on mT5 (Xue et al., 2021), a variant of T5 that was pre-trained on a multilingual dataset.

Wikilarge Zero-shot Transfer We finetuned mT5 using Wikilarge (Zhang and Lapata, 2017) dataset, and then applied the model to conduct the Chinese SS task. This experiment attempts to transfer knowledge from rich-resource language to low-resource language, leveraging the powerful cross-lingual transfer ability of mT5.

LCSTS Zero-shot Transfer LCSTS (Hu et al., 2015) is a Chinese short text summarization dataset. The tasks of sentence simplification and summarization both need the ability of compression. We trained mT5 with LCSTS and tested it on CSS, attempting to transfer knowledge from a similar task.

We also report the results of **few-shot transfer**, which continues to finetune above models with the additional dataset we have described in Section 3.5. A **few-shot baseline** that is directly finetuned with the same additional dataset but without any additional training is provided to compare with these few-shot models.

Please refer to Appendix C for the training details of the above models.

5 Evaluation Results

5.1 Automatic Evaluation Results

We use SARI and BLEU, standard metrics that were widely used in previous English sentence simplification work, to evaluate our Chinese models.

SARI The most commonly used automatic evaluation metric for sentence simplification is the SARI (Xu et al., 2016) metric, which compares the output of simplification systems with the original sentence and gold references to measure the simplicity gain. The correlation of SARI with human judgments of simplicity proved to be high (Sulem et al., 2018a). We compute SARI with the *EASSE* simplification evaluation suite (Alva-Manchego et al., 2019).[9] In our experiments, We report SARI at both the character level and word level, which means two different tokenize ways of processing Chinese text.

BLEU BLEU (Papineni et al., 2002) is a metric to measure the similarity between the system outputs and the human references, which relies on the number of n-grams in the output that match n-grams in the references, independently of position. We calculate BLEU using *NLTK* (Bird, 2006).

Previous work used FKGL (Kincaid et al., 1975) to measure the readability of the text. FKGL was tailored to be used in English only, and we do not report it in our experiments. A metric that can measure the readability of Chinese text is urgently needed in the research of Chinese sentence simplification.

Table 5 shows the automatic evaluation results. The few-shot baseline exhibits surprising results, even surpassing other few-shot models. According to the results, the data of the English SS task and short text summarization task failed to provide additional improvement for few-shot Chinese SS. We, same with Vu et al. (2022), have observed severe catastrophic forgetting when we perform cross-lingual transfer with model tuning. Perhaps model tuning is not a good option for zero- and few-shot cross-lingual transfer. Through the previous data analysis in Section 3.3, we can see that the ability of compression is not particularly important for Chinese SS, and adapting to the summarization task in advance even can harm the performance. According to the ablation experiments in Lu et al. (2021), building the pseudo corpus without FKGL

[9] We use the latest version of SARI implemented in *EASSE* which fixes bugs and inconsistencies from the traditional implementation.

	CSS		
	SARI$_{char}$	SARI$_{word}$	BLEU
Baselines and Gold Reference			
Identity Baseline	29.08	27.61	88.77
Truncation Baseline	32.95	33.18	76.36
Gold Reference	46.72	45.71	65.31
Unsupervised Method			
Lu et al. (2021)	36.27	33.39	63.47
Translate Training	36.02	34.44	71.41
Zero- and Few-shot Transfer			
Wikilarge Zero-shot Transfer	35.38	33.92	72.00
LCSTS Zero-shot Transfer	22.34	20.04	20.77
Few-shot Baseline	**37.57**	**35.97**	**74.71**
Wikilarge Few-shot Transfer	35.59	34.10	72.13
LCSTS Few-shot Transfer	34.23	32.27	64.08

Table 5: The automatic evaluation results on CSS. We use **Bold** to mark the best result and underline the second-best result. SARI$_{char}$ means the value of SARI at character level, and SARI$_{word}$ means the value of SARI at word level.

	Fluency	Meaning	Simplicity
Gold Reference	4.83	4.71	3.88
Lu et al. (2021)	4.43	4.06	2.76
Translate Training	4.67	4.49	1.59
Few-shot Baseline	4.62	4.67	1.66

Table 6: The result of human evaluation.

selector can severely harm the performance of the model. This conclusion can explain why the unsupervised methods perform worse than expected.

5.2 Human Evaluation Results

We only chose three models to manually rate their outputs due to the high cost of human evaluation, which were selected based on the SARI metric on CSS. We performed human evaluation according to the method described in Section 3.4. To maintain consistency, we chose the same 100 original sentences from CSS that were randomly selected for evaluating the quality of the dataset in Section 3.4. The human evaluation results are shown in Table 6.

The few-shot baseline obtains the best result on the SARI score and the BLEU score. However, its simplicity score is only 1.66. The few-shot baseline can not be trained adequately with a small-scale dataset. In that scenario, the model tends to replicate original sentences only with very few modifications. Therefore, this model obtains a high score both on fluency and meaning. The Translate Training model gets the lowest simplicity score, demonstrating that machine translation systems fail to bridge the gap between Chinese and English SS tasks. The human evaluation results show that we can not assess the performance of Chinese SS systems only by the value of SARI metric.

5.3 Correlation of Automatic Metrics with Human Ratings

We computed instance-level Spearman's rank correlation coefficient (Zwillinger and Kokoska, 1999) between the automatic metrics (BLEU and SARI) and human ratings, using CSS as references. Results are reported in Table 7.

The SARI$_{char}$ metric has the highest correlation with the simplicity indicator, surpassing both BLEU and SARI$_{word}$. SARI$_{word}$ also shows a moderate positive correlation with simplicity. In terms of fluency and meaning, correlations are positive but low for both SARI metrics. BLEU is dubious to apply in the evaluation of SS (Xu et al., 2016; Sulem et al., 2018b), but also shows a low positive correlation with simplicity judgments when using CSS as references. In brief, the SARI metric is not very reliable but can still be used for reference when evaluating Chinese SS, and we need more metrics to reflect the quality of the outputs in different aspects.

Metric	Fluency	Meaning	Simplicity
SARI$_{char}$	0.17	0.25	0.46
SARI$_{word}$	0.18	0.26	0.41
BLEU	0.31	0.33	0.30

Table 7: Spearman's rank correlation of human ratings with automatic metrics.

6 Chinese Sentence Simplification via Large Language Models

Large Language Models (LLMs) have demonstrated incredible ability in many natural language generation tasks and real-world applications. Recent research on sentence simplification has shown

	CSS		
	SARI$_{char}$	SARI$_{word}$	BLEU
GPT-3.5-turbo-0301			
Zero-shot	31.95	28.92	42.22
Few-shot	**39.32**	**36.57**	60.67
Vicuna-13B			
Zero-shot	23.14	20.67	23.16
Few-shot	28.68	26.56	38.04
ChatGLM-6B			
Zero-shot	35.17	32.69	56.59
Few-shot	37.74	35.70	**66.37**

Table 8: The automatic evaluation results of LLMs on CSS. We use **Bold** to mark the best result.

LLMs outperform state-of-the-art sentence simplification methods and generalize well across various low-resource languages (Feng et al., 2023). In this section, we select some representative large models and conduct zero-/few-shot experiments to evaluate them on CSS. We hope these results can supplement previous research on the cross-lingual SS capability of LLMs and serve as baselines for future studies.

According to the experiment result of Sun et al. (2023), LLaMA seems unable to understand the prompt for simplifying sentences. Therefore, we choose those LLMs that can follow Chinese instructions after instruction tuning.

GPT-3.5-turbo-0301[10] Snapshot of GPT-3.5-turbo from March 1st 2023. Unlike GPT-3.5-turbo, this model will not receive updates. We choose this stable version to ensure reproducibility.

Vicuna-13B[11] an open-source LLM trained by fine-tuning LLaMA (Touvron et al., 2023) on 70K user-shared conversations collected from ShareGPT.

ChatGLM-6B[12] an open-source LLM based on General Language Model (GLM) framework (Du et al., 2022), follows the training process similar to ChatGPT, optimized for Chinese QA and dialogue.

Then, we prompt LLMs to perform the Chinese SS task with zero-/few-shot style. In the few-shot setting, we randomly choose 5 original-simplified sentence pairs from the additional dataset as demonstrations. Please refer to Appendix D for our zero-/few-shot SS prompt template.

6.1 Analysis and Discussion

Table 8 shows the automatic evaluation result of LLMs.

Few-shot Baselines achieve better performance than Zero-shot Baselines for each LLM, which conform with our expectations. And ChatGLM performs better than Vicuna with only half the parameters. This is probably because Vicuna has not been trained specifically on the Chinese instruction dataset, although it showed Chinese capability to some extent. In fact, we even find some English characters in the simplification results of Vicuna.

It is slightly strange that GPT-3.5-turbo performs worse than ChatGLM in the zero-shot setting. There may exist a misalignment between GPT-3.5-turbo and human annotators on how to simplify Chinese sentences, and ChatGLM aligns well with humans through additional optimization for Chinese. However, GPT-3.5-turbo surpasses ChatGLM in the few-shot setting with powerful in-context learning ability, and outperforms all of baselines we report in Table 5.

The above results illustrate that LLMs like GPT-3.5-turbo can serve as a high-quality Chinese sentence simplification system.

7 Conclusion

In this paper, we are committed to facilitating research on Chinese sentence simplification. We introduced CSS, a new dataset for the evaluation of Chinese SS models. Simplifications in CSS were manually written by human annotators, and the simplification operations are also labeled. We give an in-depth analysis of CSS and reveal its characteristics. We further develop several unsupervised and zero/few-short methods on the dataset, and the experiment results can benefit future research.

[10] https://platform.openai.com/docs/models/gpt-3-5

[11] https://github.com/lm-sys/FastChat

[12] https://github.com/THUDM/ChatGLM-6B

Limitations

We only create a high-quality dataset for evaluation and a small-scale dataset for few-shot learning, since the lack of a large-scale Chinese parallel SS corpus. The available research methods for Chinese SS are limited to unsupervised learning and few-shot learning. We hope a large-scale Chinese parallel SS corpus can be released in the future. Then we can directly train more supervised models for Chinese SS.

Furthermore, we only analyze whether the current standard metrics are suitable for the evaluation of Chinese SS, and leave the work of proposing a new metric for future study. Due to time constraints, we do not perform a human evaluation for the output of LLMs. We hope to conduct a more comprehensive evaluation for LLMs in the future.

Ethics Statement

We choose original sentences from the PFR corpus, which has been released to the public. There are no intellectual property disputes for our data source. All simplifications are collected from workers we employ, and adhere to the relevant code of ethics. We pay annotators a fair salary that matches their workload.

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFF0901502), National Science Foundation of China (No.62161160339), State Key Laboratory of Media Convergence Production Technology and Systems and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. Easse: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and Mattia A Di Gangi. 2019. Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 37–44.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of german. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3302–3311.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Anna Dmitrieva and Jörg Tiedemann. 2021. A multi-task learning approach to text simplification. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 78–89. Springer.

Anna Dmitrieva, Jörg Tiedemann, et al. 2021. Creating an aligned russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. ACL Anthology.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. Editnts: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402.

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.

Richard Evans, Constantin Orăsan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 131–140, Gothenburg, Sweden. Association for Computational Linguistics.

Lijun Feng. 2008. Text simplification: A survey. *The City University of New York, Technical Report*.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *arXiv preprint arXiv:2302.11957*.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C Ziegler. 2020. Alector: A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Language Resources and Evaluation for Language Technologies (LREC)*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972.

Tomoyuki Kajiwara and M Komachi. 2018. Text simplification without simplified corpora. *The Journal of Natural Language Processing*, 25:223–249.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73.

Akihiro Katsuta and Kazuhide Yamamoto. 2019. Improving text simplification by corpus expansion with unsupervised learning. In *2019 International Conference on Asian Language Processing (IALP)*, pages 216–221. IEEE.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Dawei Li, Yanran Li, Jiayi Zhang, Ke Li, Chen Wei, Jianwei Cui, and Bin Wang. 2022. C3kg: A chinese commonsense conversation knowledge graph. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1369–1383.

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126.

Louis Martin, Éric Villemonte De La Clergerie, Benoît Sagot, and Antoine Bordes. 2020a. Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4689–4698.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020b. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text simplification by tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25.

Gustavo Paetzold and Lucia Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*. Citeseer.

Luz Rello, Clara Bayarri, Azuki Górriz, Ricardo Baeza-Yates, Saurabh Gupta, Gaurang Kanvinde, Horacio Saggion, Stefan Bott, Roberto Carlini, and Vasile Topac. 2013. Dyswebxia 2.0! more accessible text for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–2.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. Rusimplesenteval-2021 shared task: evaluating sentence simplification for russian. In *Proceedings of the International Conference "Dialogue*, pages 607–617.

Sanja Štajner, Iacer Calixto, and Horacio Saggion. 2015. Automatic text simplification for spanish: Comparative evaluation of various simplification strategies. In *Proceedings of the international conference recent advances in natural language processing*, pages 618–626.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. Bleu is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. In *CLiC-it/EVALITA*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. *arXiv preprint arXiv:2205.12647*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.

A Data source and Preprocessing

Data Source The PFR corpus (2014 version)[13] was released by Peking University, including one year's (2014) newspaper material published by the People's Daily. We chose the PFR corpus as the source of original sentences because People's Daily is the most authoritative and largest circulation newspaper in China, covering all aspects of social life.

Preprocessing The content of the corpus included the POS tag, and we restored articles to their original format and cut the text into sentences by punctuation. Then, we filtered out sentences with less than 30 tokens, since a short sentence may be difficult to simplify more.

B Simplification Examples

Original	随着深度学习、人工智能领域崛起，人类对于算力要求越来越高，显卡要想满足这些需求，必须不断迭代，但是随着摩尔定律失效，要想获得更高性能就需要更高的成本。 With the rise of deep learning and artificial intelligence, human beings are demanding more and more computing power, and graphics cards must continue to iterate for meeting these requirements, but with the failure of Moore's Law, to get higher performance will require higher costs.
Reference	随着深度学习、人工智能领域崛起，人类需要更高算力的显卡，但是这需要更高的成本。 With the rise of deep learning and artificial intelligence, humans need graphics cards with higher computing power, but this requires higher costs.
Operations	Compression; Sentence paraphrasing
Original	日侵略野心引起欧美警惕，在国际压力下日被迫放弃山东权益，从西伯利亚退兵。 Japanese aggressive ambitions caused alarm in Europe and America, and under international pressure Japan was forced to give up its rights in Shandong and retreat from Siberia.
Reference	日侵略野心引起欧美警惕。日本在国际压力下放弃山东权益，从西伯利亚退兵。 Japanese aggressive ambitions caused alarm in Europe and America. Japan gave up its rights in Shandong under international pressure and retreated from Siberia.
Operations	Compression; Sentence splitting

Table 9: Simplification examples with corresponding translations and operation tags in CSS.

C Detailed Training Settings

We train all models in PyTorch (Paszke et al., 2019), and use the HuggingFace[14] (Wolf et al., 2020) implementation of mT5 (Xue et al., 2021). We train each model on an A40 GPU. In all the experiments, we use the base version of mT5, and finetune it using Adam (Kingma and Ba, 2014) optimizer with a batchsize of 32 and a warmup ratio of 0.1. Table 10 shows the detailed setting for our models. Besides, all few-shot transfer models are trained with the setting of **few-shot baseline** in the base of zero-shot transfer models.

	Learning rate	Max-epoch	Eval steps
Lu et al. (2021)	2e-4	5	800
Translate training	2e-4	5	800
Zero-shot Transfer model	2e-4	5	800
Few-shot Baseline	1e-5	40	4

Table 10: Detailed hyperparameters of our model. Zero-shot Transfer model means **Wikilarge Zero-shot Transfer** and **LCSTS Zero-shot Transfer** in Table 5.

[13] The PFR corpus are from https://www.heywhale.com/mw/dataset

[14] https://github.com/huggingface/

D Prompt Template for Chinese Sentence Simplification

Zero-shot template

请在保留原意的基础上简化以下句子：
原句：{Original Sentence}
简化句：{Outputs}

Few-shot template

请在保留原意的基础上简化句子,以下是五个句子简化的示例：
原句：{Original Sentence}
简化句：{Simplified Sentence}

……

原句：{Original Sentence}
简化句：{Outputs}

Figure 2: The prompt template for zero-/few-shot Chinese sentence simplification

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3.1

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
we will release it in our repository after this paper is published

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3.1 & Limitations

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
There is no formal ethics committee in our institution, but our plan was discussed internally. Our data collection adheres to the relevant code of ethics.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3.1