MCTS: A Multi-Reference Chinese Text Simplification Dataset

Ruining Chong[12], Luming Lu[12], Liner Yang[12†], Jinran Nie[12], Shuhan Zhou[1], Yaoxin Li[1], Erhong Yang[12]
[1]National Language Resources Monitoring and Research Center for Print Media, Beijing Language and Culture University, China
[2]School of Information Science, Beijing Language and Culture University, China

Abstract

Text simplification aims to make the text easier to understand by applying rewriting transformations. There has been very little research on Chinese text simplification for a long time. The lack of generic evaluation data is an essential reason for this phenomenon. In this paper, we introduce MCTS, a multi-reference Chinese text simplification dataset. We describe the annotation process of the dataset and provide a detailed analysis of it. Furthermore, we evaluate the performance of some unsupervised methods and advanced large language models. We hope to build a basic understanding of Chinese text simplification through the foundational work and provide references for future research. We release our data at https://github.com/blcuicall/mcts.

1 Introduction

The task of text simplification aims to make the text easier to understand by performing multiple rewriting transformations. It can provide reading assistance for children (Kajiwara et al., 2013), non-native speakers (Paetzold, 2016) and people with language disorders (Carroll et al., 1998; Paetzold, 2016; Evans et al., 2014). Moreover, text simplification can also be used as a method of data augmentation to benefit downstream natural language processing (NLP) tasks (Van et al., 2021).

For a long time, the research of text simplification systems mainly depends on large-scale parallel corpora for training, such as WikiLarge (Zhang and Lapata, 2017) and Newsela (Xu et al., 2015). But due to the limitation of existing data in language and domain, recent work on text simplification systems has started to focus on unsupervised methods and achieves good results (Surya et al., 2019; Kumar et al., 2020; Martin et al., 2022), which makes it possible to build Chinese text simplification systems independent of large-scale parallel corpora. In this case, how to evaluate the Chinese text simplification systems becomes a problem to be solved. On the other hand, large language models have the ability to solve various NLP tasks (Thoppilan et al., 2022; Chowdhery et al., 2022). Recently a series of large language models represented by ChatGPT [1] performs well on many tasks (Qin et al., 2023; Jiao et al., 2023; Bang et al., 2023). In English text simplification, Feng et al. (2023) find that large language models outperform state-of-the-art methods and are judged to be on par with human annotators. Nevertheless, whether these models can achieve the same excellent results in Chinese text simplification remains unclear.

To solve these problems, in this paper, we introduce MCTS, a multi-reference dataset for evaluating Chinese text simplification models. MCTS consists of 3,615 human simplifications associated with 723 original sentences selected from the Penn Chinese Treebank (Xue et al., 2005) (5 simplifications per original sentence). We hope to use this dataset to measure the development status of Chinese text simplification and provide references for future research.

We design several simple unsupervised Chinese text simplification methods and test them on our proposed dataset. These methods can be served as the baselines for future studies. Furthermore, we evaluate the Chinese text simplification ability of the most advanced large language models, GPT-3.5 and ChatGPT. The results show that these large language models could outperform the unsupervised methods we set up. However, compared to human written simplification, there is still a certain gap. In summary, our contributions are listed below:

- We manually annotated a dataset that can be used for the evaluation of Chinese text simplification. It is a multi-reference dataset and contains multiple types of rewriting transfor-

†Corresponding author: Liner Yang

[1]https://chat.openai.com/chat

mations.

- We provide several text features and conducted a detailed analysis of the dataset, which could help to understand the characteristics of human Chinese text simplification.
- On the proposed dataset, we evaluated the performance of some unsupervised methods and large language models, which could serve as the baselines for future research.

2 Related Work

2.1 Evaluation Data for English text simplification

Early evaluation data for English text simplification mainly consist of sentence pairs obtained from English Wikipedia and Simple English Wikipedia through automatic sentence alignment. However, the Simple English Wikipedia was found to contain a large proportion of inadequate or inaccurate simplifications (Yasseri et al., 2012; Xu et al., 2015). And it is problematic to evaluate simplification systems with only a single reference because there are several ways of simplifying a sentence.

For the above reasons, Xu et al. (2016) introduced TurkCorpus, a multi-reference dataset for the evaluation of English text simplification. They first collected 2,359 original sentences from English Wikipedia and then obtained 8 manual reference references for every original sentence via crowdsourcing. The dataset can be used for evaluation metrics requiring multiple references, such as BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016). However, the rewriting transformations involved in TurkCorpus are very simple. Annotators were asked to simplify a sentence mainly by lexical paraphrasing but without deleting content or splitting the sentences. While another multi-reference dataset for English text simplification, HSplit (Sulem et al., 2018), only contains the rewriting transformations of sentence split, which uses the same original sentences in the test set of TurkCorpus.

In order to involve multiple transformations, Alva-Manchego et al. (2020) created the ASSET dataset. Using the same origin sentences, They extended TurkCorpus through crowdsourcing. The dataset includes rewriting transformations of lexical paraphrasing (lexical simplification and reordering), sentence splitting, and compression (deleting unimportant information). ASSET now has been adopted as a standard dataset for evaluating English text simplification systems.

Similar to ASSET, MCTS is a dataset with multiple references and multiple rewriting transformations. To our best knowledge, it is the first multi-reference dataset used for Chinese text simplification evaluation.

2.2 Unsupervised Text Simplification

Unsupervised text simplification methods do not require aligned complex-simple sentence pairs. Sai Surya et al. (2019) first attempted to realize an unsupervised neural text simplification system by importing adversarial and denoising auxiliary losses. They collected two separate sets of complex and simple sentences extracted from a parallel Wikipedia corpus and trained on them with auto-encoders. Lu et al. (2021) found that during the process of neural machine translation, it is possible to generate more high-frequency tokens. According to this finding, they built a pseudo text simplification corpus by taking the pair of the source sentences of the translation corpus and the translations of their references in a bridge language, which could be used to train text simplification models in a Seq2Seq way. Martin et al. (2022) leveraged paraphrase data mined from Common Crawl and used ACCESS (Martin et al., 2020), a method to make any sequence-to-sequence model controllable, to generate simplifications and not paraphrases at test time. Their method achieved good results and was considered the state-out-of-art unsupervised text simplification method.

2.3 Large Language Models

Compared to general pre-trained models, large language models are also typically based on the transformer architecture but are much larger in scale, such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022) and OPT (Zhang et al., 2022). They can handle various NLP tasks through the given instructions, which do not require any gradient updates (Brown et al., 2020).

ChatGPT is obtained by fine-tuning a GPT-3.5 via reinforcement learning from human feedback (RLHF) (Christiano et al., 2017). As a large language model for intelligent human-computer dialogue, it can answer user input with high quality. ChatGPT has recently attracted significant attention from the NLP community, and there have been many studies on it (Qin et al., 2023; Guo et al., 2023; Yang et al., 2023). However, exploring these

Original	为适应大西南进出口物资迅速增长的需要，北部湾沿海四市开始了新一轮建港热潮。 In order to adapt to the rapid growth of import and export materials in the southwest, four coastal cities in the Beibu Gulf have started a new wave of port construction.
Reference	大西南进口和出口物资的需要迅速增长，北部湾沿海四座城市兴起了新一轮港口建设。 The demand for imported and exported materials in the southwest has grown rapidly. Four coastal cities in the Beibu Gulf have begun a new round of port construction.
Original	中国又一条煤炭运输大通道——连接天津蓟县与天津港之间的蓟港铁路日前破土动工。 Another major coal transportation corridor in China - the Jigang Railway connecting Tianjin Jixian County and Tianjin Port - has recently broken ground.
Reference	蓟港铁路连接天津蓟县和天津港，用于煤炭运输，前几天开始建造。 Jigang Railway connects Tianjin Jixian County and Tianjin Port for coal transportation, and construction began a few days ago.
Original	按设计，“进步M-24”号载重飞船可同轨道站在无人操纵的情况下进行自动对接。 According to the design, the "Progress M-24" heavy-duty spacecraft can automatically dock with the orbital station under unmanned control.
Reference	按照设计，无人操控的“进步M-24”号飞船可以自动对接轨道站。 According to the design, the unmanned "Progress M-24" spacecraft can automatically dock with the orbital station.

Table 1: Examples of simplifications collected for MCTS

models in Chinese text simplification is still lacking.

3 Creating MCTS

In this section, we describe more details about MCTS. In section 3.1, we introduce the preparation of original sentences. And in section 3.2, we introduce the annotation process of MCTS.

3.1 Data Preparation

We use Penn Chinese Treebank (CTB) as the source of the original sentence in the dataset. CTB is a phrase structure tree bank built by the University of Pennsylvania. It includes Xinhua news agency reports, government documents, news magazines, broadcasts, interviews, online news, and logs. We first filtered out the simple sentences using a filter based on the average lexical difficulty level in HSK to ensure that the original sentences we choose are sufficiently complex. Then we manually selected from the remaining sentences. Finally, we obtained 723 news sentences as the original sentence.

3.2 Annotation Process

MCTS is an evaluation dataset that is completely manually annotated. The detailed annotating process is as follows.

Annotator Recruitment All the annotators we recruited are native Chinese speakers and are undergraduate or graduate students in school. Most of them have a background in linguistics or computer science. All annotators needed to attend a training course and take the corresponding Qualification Test (see more details below) designed for our task. Only those who have passed the Qualification Test could enter the Annotation Round.

Simplification Instructions We provided the exact instructions for annotators for the Qualification Test and the Annotation Round. In the instructions, we defined three types of rewriting transformations.

- Paraphrasing: Replacing complex words or phrases with simple formulations.
- Compression: Deleting repetitive or unimportant information from the sentence.
- Structure Changing: Modifying complex sentence structures into simple forms.

Compared to rewriting transformations involved in ASSET, we replaced sentence splitting with structural changing. The latter covers a broader range and is more consistent with the actual situation of simplifying Chinese sentences. Besides, the paraphrasing transformation in Chinese is much more flexible than in English. It includes not only the substitution of synonyms but also the interpretation of complex phrases or idioms. For every rewriting transformation, we provided several examples. Annotators could decide for themselves which types of rewriting to execute in any given original sentence.

Qualification Test At this stage, we provided 20 sentences to be simplified. Annotators needed to simplify these sentences according to the instructions given. We checked all submissions to filter

out annotators who could not perform the task correctly. Of the 73 people who initially registered, only 35 passed the Qualification Test (48%) and worked on the task.

Annotation Round Annotators who passed the Qualification Test had access to this round. To facilitate annotating work, we provided a platform that can display the difficulty level of words in a text. We collected 5 simplifications for each of the 723 original sentences. Table 1 presents a few examples of simplifications in MCTS, together with their English translation.

4 Dataset Analysis

Following ASSET (Alva-Manchego et al., 2020), we report a series of text features in MCTS and study the simplifications in the dataset through them.

4.1 Text Features

We calculated several low-level features for all simplification examples to measure the rewriting transformations included in MCTS. These features are listed below.

- Number of sentence splits: The difference between the number of sentences in the simplification and 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: The character-level Levenshtein distance (Levenshtein et al., 1966) for replace operations only divided by the length of the shorter string in the original sentence and simplification. As described in ASSET, ignoring insertions and deletions can make this feature independent of compression level and serve as a proxy for measuring the lexical paraphrases of the simplification.
- Proportion of words deleted, added and reordered: The number of words deleted/reordered 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.
- Lexical complexity score ratio: We compute the score as the mean squared lexical difficulty level in HSK. The ratio is then the value of this score on the simplification divided by that of the original sentence, which can be considered as an indicator of lexical simplification.
- Dependency tree depth ratio: The ratio of the depth of the dependency parse tree of the simplification relative to that of the original sentence. Follwing ASSET (Alva-Manchego et al., 2020), we perform parsing using spaCy [2]. This feature can reflect structural simplicity to a certain extent.

4.2 Results and Analysis

The density of all these features is shown in Figure 1. We can see that sentence splitting operation appears not frequently on MCTS. By observing the data, we believe that this is due to the characteristics of the Chinese. Compound sentences are commonly used in Chinese and one sentence consists of two or more independent clauses. During the simplification, annotators tend to rewrite a complex sentence with nested clauses into compound sentences rather than multiple simple sentences. So this is not to say that Chinese text simplification rarely involves sentence structure change, but that the way of structural change is not limited to sentence splitting.

Although we introduced compression as a rewriting transformation in the simplification instructions, the compression ratio is not too concentrated on the side less than 1.0. The reason is that, on the one hand, the annotators tend to retain as much semantic information as possible, and on the other hand, more characters may be added when paraphrasing.

By analyzing replace-only Levenshtein distance, we can see that the simplifications in MCTS have a considerable degree of paraphrasing the input as simplifications are distributed at all levels. Regarding the distribution of deleted, added, and reordered words, we can find that the peaks all occur at positions greater than 0.0. This further reveals the plentiful rewriting operations contained in MCTS.

In terms of lexical complexity, we can clearly see the high density of ratios less than 1.0, indicating that simplification has significantly lower lexical complexity compared to the original sentence. Some instances have a lexical complexity

[2] https://github.com/explosion/spaCy

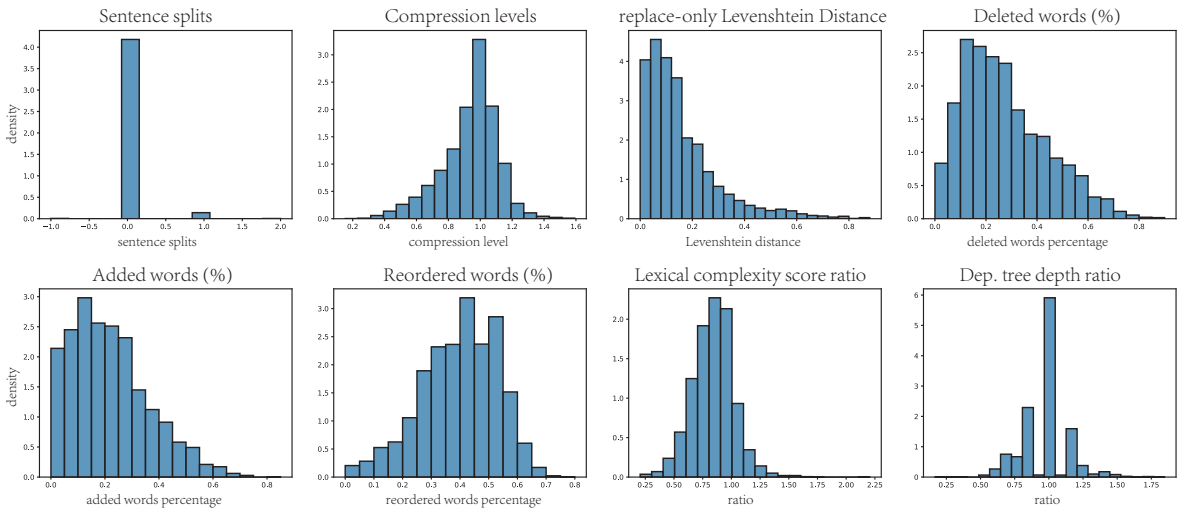

Figure 1: Density of text features in simplifications from MCTS

ratio greater than 1.0, which may be due to deleted simple words in the process of sentence compression.

Finally, the dataset shows a high density of a 1.0 ratio in dependency tree depth. This may indicate that significant structural changes were not made.

5 Experiment

In order to measure the development status of Chinese text simplification and provide references for future research, we conducted a series of experiments on the proposed MCTS.

5.1 Methods

We attempt several unsupervised Chinese text simplification methods and large language models and provided their results on MCTS. The first three are unsupervised methods that utilize automatic machine translation technology. We use Google Translator [3] to translate. These unsupervised methods can be used as the baselines for future work.

Direct Back Translation As high-frequency words tend to be generated in the process of neural machine translation (Lu et al., 2021), back translation is a potential unsupervised text simplification method. We translated the original Chinese sentences into English and then translated them back to obtain simplified results. We chose English as the bridge language because of the rich bilingual translation resources between Chinese and English.

Translated Wiki-Large Translating existing text simplification data into Chinese is a simple way to construct pseudo data. We translated English sentence pairs in Wiki-Large into Chinese sentence pairs and used them to train a BART-based (Lewis et al., 2020) model as one of our baselines.

Cross-Lingual Pseudo Data In addition to the above two methods, we also designed a simple way to construct pseudo data for Chinese text simplification, which can leverage the knowledge from English text simplification models. As shown in Figure 2, we first collect a large amount of Chinese sentence data, for example, the People's Daily Corpus. Then, we translate these sentences into English and simplify them using existing English text simplification models. Finally, we translate the simplified English sentences back into Chinese and align them with the original Chinese sentences to obtain parallel data. To ensure data quality, we filter the obtained parallel data from three aspects: simplicity, fluency, and semantic retention. For simplicity, we calculate the average lexical difficulty level for both the original sentence and the simplified sentence. Only when the difficulty level of the simplified sentence is significantly reduced compared to the original sentence, this parallel sentence pair will be retained. For fluency, we calculate the perplexity for the simplified sentences and filter out sentences above the preset threshold. For semantic retention, we use sentence-transformers toolkit (Reimers and Gurevych, 2020) to calculate the semantic similarity between the original sentence and simplified sentence, and also filter out sentences that exceed the preset threshold. Using the filtered data, we train a BART-base model.

[3] https://translate.google.com/

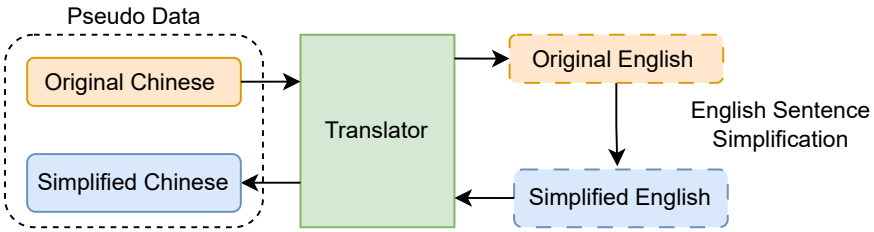

Figure 2: Pseudo data acquisition process

Large Language Models We chose two advanced large language models to conduct experiments: *gpt-3.5-turbo* and *text-davinci-003*. Both of them are based on GPT-3.5. The former is the most capable GPT-3.5 model and is optimized for chatting. The latter is the previous model, which can execute any language task according to instructions. We translated the simplification prompt used by Feng et al. (2023) as our prompt. More details about the prompt can be found in Table 2. The experiment was conducted under the zero-shot setting.

Our Prompt
我想让你把我的复杂句子替换成简单的句子。你要保持句意不变，但使句子更简单。 复杂句：{Complex Sentence} 简单句：{Simplified Sentence(s)}

Table 2: Prompt for Chinese text simplification

5.2 Automatic Metrics

Following previous work, we choose three metrics for evaluation: SARI (Xu et al., 2016), BLEU (Papineni et al., 2002) and HSK Level (Kong et al., 2022).

SARI SARI (Xu et al., 2016) is a commonly used evaluation metric for text simplification. Comparing system outputs to multiple simplification references and the original sentences, SARI calculates the mean of the n-gram F1 scores of *add*, *keep*, and *delete*. In our experiment, we tokenize sentences using Stanford CoreNLP[4] and use the EASSE toolkit [5] (Alva-Manchego et al., 2019) to calculate SARI.

BLEU BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) was initially used to evaluate the quality of machine translation. By calculating the N-gram and counting the times that can be matched, BLEU can reflect the closeness between system outputs and references. Just like calculating SARI, we use the EASSE toolkit to calculate the BLEU score.

HSK Level In order to measure the complexity of Chinese sentences, we import HSK Level. HSK is the Chinese proficiency test designed for non-native speakers [6]. It provides a vocabulary of nine levels from easy to difficult. Following previous work (Kong et al., 2022), we count the proportion of words at levels 1-3 and 7+ in system outputs. The higher the proportion of words in levels 1-3 (7+), the easier (more challenging) the outputs are understood. Our specific implementation of this metric is the same as Kong et al. (2022).

5.3 Human Evaluation

In order to obtain more comprehensive evaluation results, we further conduct human evaluation. Following the previous work (Dong et al., 2019; Kumar et al., 2020), we evaluate the Chinese text simplification systems on three dimensions:

- Fluency: Is the output grammatical?
- Adequacy: How much meaning from the original sentence is preserved?
- Simplicity: Is the output simpler than the original sentence?

We provide simplifications generated by different systems for the recruited volunteers. And we ask the volunteers to fill out a five-point Likert scale (1 is the worst, 5 is the best) about these simplifications for each dimension. Additionally, following Feng et al.'s work (2023), we measure the volunteers' subjective choices by ranking the simplifications to focus on actual usage rather than evaluation criteria.

[4] https://github.com/stanfordnlp/CoreNLP
[5] https://github.com/feralvam/easse
[6] https://www.chinesetest.cn

Method	SARI ↑	BLEU ↑	L1-3 (%) ↑	L7+ (%) ↓
Source	22.37	84.75	40.24	44.90
Gold Reference	48.11	61.62	46.25	39.50
Direct Back Translation	<u>40.37</u>	48.72	39.19	45.44
Translated Wiki-Large	28.30	**82.20**	40.32	44.92
Cross-Lingual Pseudo Data	38.49	<u>63.06</u>	<u>41.57</u>	<u>44.24</u>
gpt-3.5-turbo	**42.39**	49.22	**43.68**	**41.29**
text-davinci-003	37.97	36.18	38.80	45.32

Table 3: The automatic evaluation results on the test set of MCTS. ↑ The higher, the better. ↓ The lower, the better. **Bold** means the best result, and <u>underline</u> means the second-best result.

6 Results

We divide all the 723 sentences in MCTS into two subsets: 366 for validation and 357 for testing the Chinese text simplification models. In this section, we report the evaluation results on the test set of MCTS.

6.1 Results of Automatic Evaluations

The results of automatic evaluations are shown in Table 3. In addition to the model results, we also report the score of the source and gold reference. The source scores are calculated on the unedited original sentence. And we calculate the gold reference scores by evaluating each reference against all others in a leave-one-out scenario and then averaging the scores.

To our surprise, direct back translation gets the best SARI score among the unsupervised methods. But regarding HSK level, the performance of direct back translation is not good, even worse than the source. We find that many rewrite operations were generated during the back translation process, which is highly correlated with the SARI score. But due to the lack of control over simplicity, direct back translation is more like a sentence paraphrase method than text simplification. This may be why it performs poorly on the HSK level.

The translated Wiki-Large method gets the best BLEU score but the lowest SARI score among all methods. In fact, the system output has hardly changed compared to the original sentence. As the unedited source gets the highest BLEU score of 84.75, we believe the single BLEU value cannot be used as an excellent indicator of text simplification. Because there is a significant overlap between the original sentence and the references. As for the poor performance of translated Wiki-Large method, we believe it is due to the large amount of noise contained in the translated training data.

The SARI score of the cross-lingual pseudo data method is 38.49, which is between the other two unsupervised methods. But it performs better on the HSK level than the other two. This may be because the model learned simplification knowledge from pseudo data that was transferred from the English text simplification model.

In terms of the large language models, the gpt-3.5-turbo significantly performs better than text-davinci-003 and it achieves the best scores on SARI and HSK levels. However, compared to the gold reference, the performance of gpt-3.5-turbo is still insufficient.

6.2 Results of Human Evaluations

We conducted human evaluations on three representative methods, namely direct back translation, cross-lingual pseudo data, and gpt-3.5-turbo. We recruited three volunteers to conduct the evaluation. All of them have a background in linguistics. We selected 30 sentences from the test set of MCTS for each volunteer and provided them with the original sentences and the outputs of these methods. For the convenience of comparison, a randomly selected reference for each sentence was additionally provided. Volunteers were asked to rate the simplification of these four groups. The results of the human evaluation are shown in Table 4.

We can see that the gold reference gets the best average score and rank. It is significantly superior to the output results of other simplification systems. For detail, it gets the best simplicity score of 4.20 and the best fluency score of 4.68. Due to some degree of sentence compression, it does not achieve the best adequacy score but only 4.31.

Method	Simplicity ↑	Fluency ↑	Adequacy ↑	Avg. ↑	Rank ↓
Gold Reference	**4.20**	**4.68**	4.31	**4.40**	**1.97**
Direct Back Translation	3.42	4.36	**4.72**	4.17	2.88
Cross-Lingual Pseudo Data	4.11	4.46	3.88	4.15	2.86
gpt-3.5-turbo	4.17	4.46	4.43	4.35	2.29

Table 4: The human evaluation results. Avg. means the average score of fluency, adequacy and simplicity. Rank means the subjective ranking for the simplifications.

As for the direct back translation method, despite its excellent performance in adequacy, it achieves the lowest simplicity score due to the lack of corresponding control measures. On the contrary, the cross-lingual pseudo data method performs well in terms of simplicity but does not perform well in terms of adequacy. Because it tends to perform more sentence compression, which removes lots of semantic information. These two unsupervised methods get a similar average score and rank score.

The gpt-3.5-turbo gets the second-best results among all metrics. By analyzing the average score and the rank score, we can find that it is significantly better than the two unsupervised simplification methods. But compared to the gold reference, there is still a certain gap. Our experiment has shown that under the zero-shot setting, there is still room for further improvement in the large language model's Chinese text simplification ability.

7 Conclusion

In this paper, we introduced the MCTS, a human-annotated dataset for the validation and evaluation of Chinese text simplification systems. It is a multi-reference dataset that contains multiple rewriting transformations. By calculating the low-level features for simplifications, we have shown the rich simplifications in MCTS, which may be of great significance for understanding the simplification and readability of Chinese text from a linguistic perspective. Furthermore, we tested the Chinese text simplification ability of some unsupervised methods and advanced large language models using the proposed dataset. We found that even advanced large language models are still inferior to human simplification under the zero-shot setting. Finally, we hope our work can motivate the development of Chinese text simplification systems and provide references for future research.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (No. 23YCX131).

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. 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: System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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. Association for the Advancement of Artificial Intelligence.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts,

Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

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.

Richard Evans, Constantin Orasan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. Association for Computational Linguistics.

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

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

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.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking framework for unsupervised simple definition generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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.

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

Louis Martin, Angela Fan, Éric Villemonte De La Clergerie, Antoine Bordes, and Benoît Sagot. 2022. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664.

Gustavo Henrique Paetzold. 2016. *Lexical simplification for non-native english speakers*. Ph.D. thesis, University of Sheffield.

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.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4512–4525.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. 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.

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.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hoang Van, Zheng Tang, and Mihai Surdeanu. 2021. How may i help you? using neural text simplification to improve downstream nlp tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4074–4080.

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.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Taha Yasseri, András Kornai, and János Kertész. 2012. A practical approach to language complexity: a wikipedia case study. *PloS one*, 7(11):e48386.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *EMNLP 2017: Conference on Empirical Methods in Natural Language Processing*, pages 584–594. Association for Computational Linguistics.