Chinese Sentence Paraphrasing

Anonymous ACL submission

Abstract

Sentence paraphrasing involves understanding the semantics and generating alternative expressions that are equivalent to the original sentence but not identical. However, there lack of an evaluation metric for paraphrasing that aligns well with human annotation and a lack of high-quality Chinese paraphrase datasets which makes it difficult to train a Chinese paraphrase model. To address these challenges, we present the first large-scale automatically constructed Chinese sentence paraphrase corpus, consisting of 9.45 million annotated sentence pairs for paraphrasing. We also introduce a core dataset with 2.5 thousand Chinese sentence pairs that are completely paraphrased by the crowd and annotated by experts. With this high-quality data, we establish an automatic evaluation metric for Chinese paraphrasing evaluation, achieving a Spearman coefficient of 0.726 in human-annotated data and significantly outperforming existing metrics. Additionally, we build a strong baseline for Chinese paraphrasing generation with few entity and logical errors while preserving the meaning of the sentence and generating diverse and innovative sentences.[1]

1 Introduction

Sentence paraphrasing (Bhagat and Hovy, 2013) aims to change expressions or improve the readability of a sentence by altering its structure and replacing words with synonyms. In machine translation, researchers find that paraphrasing the source language sentence can enhance translation quality (Thompson and Post, 2020). Additionally, in text summarization, sentence paraphrasing can assist in generating more concise and accurate summaries (Nayeem et al., 2018; Tang et al., 2023). Furthermore, sentence paraphrasing plays a crucial role in tasks such as question-answering systems (Gan and Ng, 2019) and information retrieval (Zhang et al., 2015; Ferguson et al., 2018).

In recent years, with the advancement of deep learning, methods utilizing neural network models for English sentence paraphrasing (Kumar et al., 2020; Huang and Chang, 2021) are widely applied and researched. However, sentence paraphrasing in Chinese still has two main challenges.

First of all, there is a lack of scientific and systematic automatic evaluation metrics for sentence paraphrasing, whether it is in English or Chinese. Existing works (Ormazabal et al., 2022a; Dou et al., 2022a) mostly adopt iBLEU (Sun and Zhou, 2012) as the evaluation metric for paraphrasing, while others utilize traditional text evaluation metrics such as BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020). However, paraphrasing sentence pairs exhibit high semantic similarity while having significant differences in vocabulary, phrases, and structure. This gap makes it difficult for traditional evaluation metrics to evaluate the quality of paraphrased sentences, as higher scores indicate higher semantic similarity but lower degrees of paraphrasing, thus failing to provide a comprehensive assessment of paraphrase quality. Shen et al. (2022) propose a new evaluation metric, ParaScore, for sentence paraphrasing. However, due to the low quality of the validation dataset used, whose sentence pairs marked as high score only retain the correct semantics and with a low degree of paraphrase, ParaScore has poor generalization.

Secondly, there lack of a high-quality dataset for Chinese sentence paraphrasing. Shen et al. (2022), by utilizing the English paraphrase dataset BQ-Para (Chen et al., 2018), construct pseudo-Chinese paraphrase datasets through various paraphrase generation algorithms and annotating scores. However, these paraphrase generation algorithms suffer from several issues as shown in Table 1. For example, back-translation paraphrase methods (Prabhumoye et al., 2018) may introduce semantic errors and entity misalignment, while large model-based approaches (Witteveen and Andrews, 2019) may

[1]The API will be made public after acceptance.

	Original Sentence	Paraphrased Sentence	ST	SC	SR
Back Translation	他，是冷氏的大少爷，是邪恶霸道冷酷的组合的老大，是嗜血的帮主，非常的神秘。	他是凌希少爷，霸道冷酷的邪恶集团总裁，嗜血成性的黑帮老大，神秘至极。	✓	×	✓
	“你疯了！”约塞连生气地对邓巴喊道，“你究竟为什么要这么说？”	“你疯了！”约塞连愤怒地冲邓巴吼道。“你到底为什么要这么说？”	×	✓	×
GPT-3.5 Turbo	今天上午，宁波·凉山东西部协作联席会在四川凉山召开。会上，宁波向凉山捐赠消防车150辆，总价值3750万元。	今日上午，宁波·凉山东西部协作联席会在四川凉山召开，宁波方捐赠凉山消防车150辆，总价值3750万元。	×	✓	×
	11月12日下午，澎湃新闻从应急管理部森林消防局机动支队获悉，应急管理部森林消防局机动支队张洪顺支队长带支队前指、五大队共67人向江西九江火场机动，全程130公里。	11月12日下午，应急管理部森林消防局机动支队张洪顺支队长带领67人前往江西九江火场，全程130公里。	✓	×	×
Human	走过去搂着汪珊说：“老婆，对不起，我错了，我不该这么说你。困难只是短暂的，再忍忍，孩子大一点就好了。”	走到汪珊身边，搂住她，轻声细语道：“宝贝，对不起，脱口而出的话真的太不恰当了。可是，困难也不过是短暂的，只要我们在一起陪伴孩子，他们长大就好。”	✓	✓	✓

Table 1: Limitations of different Chinese paraphrasing generation methods, where ST represents sentence transformation, SC represents semantic consistency, and SR represents synonym replacement.

exhibit few changes or omit crucial content. Most examples annotated high scores in BQ-Para (Chinese) can only guarantee semantic correctness and basic synonym substitution, with a relatively low degree of paraphrasing. Consequently, the existing Chinese paraphrase datasets suffer from low quality. This makes it more challenging to train a high-quality Chinese sentence paraphrasing model.

To address these challenges, we first establish a core Chinese paraphrase corpus $\text{CSPC}_{\text{core}}$ through human paraphrasing and expert annotation. Based on $\text{CSPC}_{\text{core}}$, we propose a neural-based comprehensive evaluation metric for automatically assessing paraphrase qualities using 8 designed feature extractors, achieving state-of-the-art correlation with human annotations on $\text{CSPC}_{\text{core}}$. Furthermore, we collect 17 million parallel translation data, 1.2 million back-translation data, and 140 thousand sentences paraphrased by GPT-3.5 Turbo and then filter them to obtain 9.45 million high-quality Chinese paraphrase sentences, a new large-scale and high-quality automatically constructed Chinese sentence paraphrase corpus $\text{CSPC}_{\text{auto}}$. Finally, through three stages of training, combined with our proposed entity-aligned tokenizer, we present a strong baseline for Chinese sentence paraphrasing.

The contributions are as follows:

• We introduce the first large-scale, high-quality Chinese paraphrase dataset, which includes 9.45 million sentences, and over 2,585 human paraphrased sentence pairs with experts annotated.

• We formulate 8 paraphrasing rules and design their corresponding feature extractors. Through feature engineering and pattern recognition, we develop a neural-based evaluation metric for Chinese sentence paraphrasing. In experiments, our metric achieves state-of-the-art performance in terms of correlation with human annotations.

• Considering the characteristics of Chinese sentence paraphrasing, we propose a model-agnostic entity-aligned training strategy. Building upon this approach, we develop a strong baseline and our proposed Chinese sentence paraphrasing models can generate diverse, high-quality sentences that meet application standards, as demonstrated by our paraphrasing evaluation metric and case studies.

2 Related Work

Sun and Zhou (2012) use statistical machine translation for paraphrase generation and propose iBLEU to evaluate the quality of the paraphrases. Witteveen and Andrews (2019) utilize pre-trained language models for paraphrase generation and evaluate the fine-tuned GPT-2 model using Rouge-L and BLEU. Ormazabal et al. (2022b) employ parallel corpora for paraphrase generation and achieve better results on iBLEU compared to round-trip machine translation.

However, these approaches have some limitations, either by using weak baselines for comparison or inappropriate evaluation metrics to assess the paraphrasing abilities of the models. Dou et al. (2022b) propose new standards for paraphrase identification and train more powerful paraphrase generation models by creating high-quality English datasets. Shen et al. (2022) point out that existing evaluation metrics for paraphrasing cannot align well with human annotations, and thus propose a new evaluation method called ParaScore. They also create the first Chinese paraphrase dataset, BQ-para (Chinese), using paraphrase algorithms. However, due to the limited capability of existing Chinese

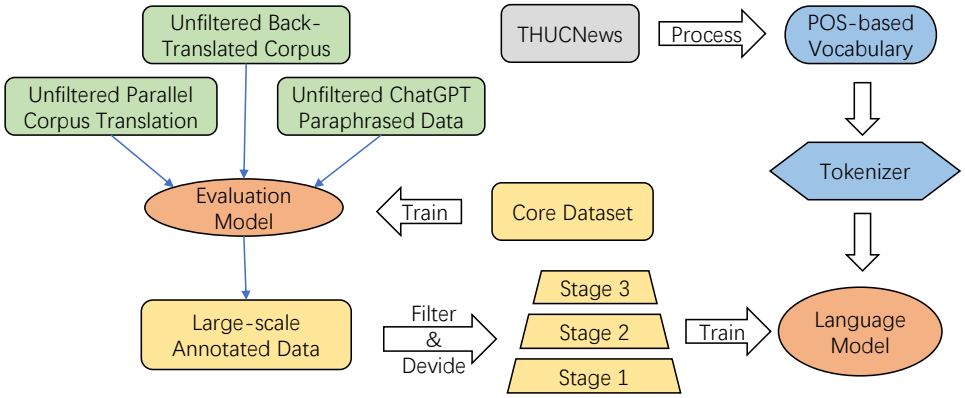

Figure 1: Pipeline of collecting Chinese paraphrase datasets, establishing paraphrase evaluation metrics, and training a paraphrase generation model.

paraphrase generation algorithms, the quality of this dataset is not high, resulting in poor generalizability of ParaScore.

Lin et al. (2020) enhance the quality of generated text using linguistic knowledge through the retrieve, locate and generate pipeline, and establish a translation-based Chinese news paraphrase as a benchmark for Chinese sentence paraphrasing. However, such data still suffer from the aforementioned quality issues, which result in the weak generalization of the generation model.

3 Dataset Construction

3.1 Core Dataset

Firstly, we create a completely human-paraphrased and expert-annotated core dataset, $\text{CSPC}_{\text{core}}$. Each crowd worker is assigned a number of Chinese sentences and they are required to paraphrase these sentences with the same semantics but different expressions ($0.04 per valid sentence). These Chinese sentences are extracted from various sources such as novels, news, and books. However, considering the differences in the Chinese language competence among the crowd workers and the practical work environment, these human-paraphrased Chinese sentences are not directly used as the dataset. Instead, they are annotated by experts to assess their quality of paraphrasing ($0.01 per sentence).

Based on the granularity of annotation, we divide them into 730 pairs of fine-annotated Chinese sentences and 1933 pairs of coarse-annotated Chinese sentences. The fine-grained annotation scores ranged from 0 to 5, with 0 indicating poor paraphrasing quality and 5 indicating high paraphrasing quality. The coarse-grained annotation scores were either 0 or 1, with 0 indicating inadequate paraphrasing and 1 indicating sufficient paraphrasing.

For the Chinese sentence paraphrasing during dataset construction, we have three basic criteria: 1) Structural Transformation: Changes in time, place, and narrative order are required; 2) Synonym Replacement: Synonyms are preferred for nouns, verbs, adjectives, and other words; 3) Semantic Consistency: The paraphrased sentence must retain all the semantics of the source sentence without any additions or deletions. We instruct the crowd workers to paraphrase the Chinese sentences according to these three basic criteria and inform the experts to score and annotate the paraphrased sentence pairs based on these criteria as well.

3.2 Large Scale Annotated Dataset

As for the large-scale annotated dataset, $\text{CSPC}_{\text{auto}}$, we start with the $\text{CSPC}_{\text{core}}$ in the previous step and propose a new automatic evaluation metric called SPScore-zh to evaluate the quality of paraphrasing (details are described in Section 4). First, we filter out Chinese sentences with less than 15 characters from the en-zh data of the United Nations Parallel Corpus. Then, we use Google Translate to translate the English part into Chinese and combined it with the original Chinese part in the corpus to form Chinese sentence pairs. We collect a total of 17 million unfiltered parallel translation data as pseudo-paraphrase sentence pairs. We then use SPScore-zh to score these data and filter 8.9 million parallel translation data as a part of the final Chinese paraphrase sentence pairs.

In addition, we utilize the m2m100 translation model (Fan et al., 2021) to translate 1.2 million Chinese sentences into Arabic and then perform back-translation using the same translation model. Furthermore, taking advantage of large language mod-

Paraphrase Prompt
请帮我改写输入的句子，要求： 1. 确保原文的语法、拼写和句式没有错误，这样可以让改写结果更加准确和流畅。 2. 在改写时，尽可能使用自己的语言来表达原文中的含义，改变句子结构，避免直接复制原文的短语和句式。 3. 对一些机构、个人、专有名词可以进行其他叫法的替换。

Table 2: Paraphrase instructions for GPT-3.5 Turbo.

els, we attempt to generate paraphrased sentences using GPT-3.5 Turbo. After various attempts, we adopt the above prompt designs as shown in Table 2, that allows GPT-3.5 to perform Chinese paraphrase generation. Using this method, we generate 140 thousand pseudo-Chinese paraphrase sentence pairs. Finally, we use the 1.2 million unfiltered back-translated corpus and the 140 thousand unfiltered GPT-3.5 paraphrase data as the original corpora, and through SPScore-zh annotation filtering, obtain 550 thousand high-scored Chinese paraphrase sentence pairs.

Among these three sources of data collection, we randomly select a few samples and find that, under the same annotation score, the translation data from parallel corpora has the lower quality, while the data generated by GPT-3.5 has the higher quality. This includes considerations of readability and logical coherence. Therefore, in Section 5.3, when dividing the dataset into three stages based on the scores, we will take into account both the scores and the data sources.

4 New Paraphrase Metric: SPScore-zh

In Section 3.1, we propose three basic criteria for paraphrasing. Now, we further expand these three basic criteria into eight computable features, making them suitable for direct computation and automatic evaluation. Finally, through feature engineering, we obtain this Chinese-specific automatic evaluation metric SPScore-zh which is highly correlated with human annotations.

The Spearman and Pearson coefficients are statistical measures used to assess the correlation between automatic evaluation metrics and human-annotated scores. The ablation studies on these features are shown in Table 3.

4.1 Structural Transformation

First, for criterion one, structural transformation, we expand it into four computable features for each Chinese sentence pair:

• Appositive Character Similarity (ACS): Comparing the proportion of characters in the original sentence and the paraphrased sentence at the same positions. By examining the degree of character overlap, this metric provides the fidelity of the paraphrase and how the meaning and structure of the original text are preserved.

• Substring Positional Alignment (SPA): Relative positions of substrings in the two sentences and evaluate the positional transformations between substrings. By examining the positional alignments of substrings, we know how the sentences are structured and how the information is organized within them. Function $\mathrm{D}_\alpha(a, b)$ means distance of two words a and b in sentence α and function $\mathrm{RD}_\alpha(a)$ means relative distance of the word a in sentence α. The input I in Equation (1) can be substrings, phrases, words, and characters. Here, the input I is shared substrings in sentences A and B:

$$\mathrm{SPA}(A,B) = \frac{1}{|I|^2}\sum_{i=0}^{|I|}\sum_{j=i}^{|I|} |\mathrm{RD}_A(I_i) - \mathrm{RD}_B(I_j)| \quad (1)$$

Metric	Core Dataset (Test Set)	
	Pearson↑	Spearman↑
w/o ACS	0.6855	0.7089
w/o SPA	0.7013	0.7246
w/o PMD	0.7040	0.7254
w/o WPA	0.7049	0.7255
w/o LIS	0.6425	0.6594
w/o SBLEU	0.6767	0.7013
w/o SL	0.7049	0.7164
w/o LR	0.7034	0.7238
SPScore-zh	**0.7059**	**0.7261**

Table 3: Ablation study about SPScore-zh on $\mathrm{CSPC}_{\mathrm{core}}$.

• Phrase Mixing Degree (PMD): Degree of phrase mixing based on comma separation. The degree of phrase mixing can provide insights into the syntactic complexity and cohesion of a sentence. Function $\mathrm{MD}(S)$ calculates the mixing degree of sentence S and for the phrase in sentences A, B and their shared phrase I:

$$\mathrm{MD}(S) = \frac{1}{|S|}\sum_{i=0}^{|I|}\sum_{j=i}^{|I|} |\mathrm{D}_S(I_i, I_j)|$$
$$\mathrm{PMD}(A,B) = \frac{1}{|I|}\,|\mathrm{MD}(A) - \mathrm{MD}(B)| \quad (2)$$

• Word Positional Alignment (WPA): Evaluating the positional transformations between words. It involves analyzing the structural correspondence between words in order to understand how their positions change or remain consistent when comparing two sentences. WPA is calculated by Equation (1) using shared words I.

These four features calculate the degree of structural transformations at different granularities of a sentence, from characters to substrings.

4.2 Synonym Replacement

Second, for criterion two, we expand it into two computable features:

• Longest Identical Substring (LIS): Finding the longest substring that is exactly the same between the original sentence and the paraphrased sentence. This can help determine whether the paraphrase maintains the structure, syntax, and sentence organization of the original text. If the longest identical substring is only a small fragment, it suggests that the paraphrase may have undergone significant changes in terms of structure or syntax.

• BLEU Score for Substrings (SBLEU): Considering only substrings with a length greater than three and using the BLEU score. By only considering substrings with a length greater than three, we can better capture the changes of key information during paraphrasing. These substrings often represent important modifiers in the sentence, and their alterations are crucial for maintaining the coherence of the sentence.

These two features measure the degree of replacement and semantic similarity at different semantic granularities, from words to substrings.

Metric	Core Dataset (Test Set)	
	Pearson↑	Spearman↑
BERTScore	0.467	0.521
BLEU	0.635	0.623
Self-iBLEU	0.649	0.665
METEOR	0.582	0.592
ROUGE-1	0.508	0.509
ROUGE-2	0.608	0.589
ROUGE-L	0.628	0.635
ParaScore	0.669	0.689
SPScore-zh	**0.706**	**0.726**

Table 4: Performance about different metrics on $\text{CSPC}_{\text{core}}$ test set. Most of the existing metrics are negatively correlated with human annotations and we compare them by taking their absolute values.

4.3 Semantic Consistency

Finally, for criterion three, we expand it into two computable features:

• Sentence Length (SL): Indicating the number of characters in the sentence that needs to be paraphrased. The length of a sentence can be used to determine the extent of information conveyance and grammatical structure.

• Length Ratio (LR): The length ratio can be used to detect whether there is any information loss or redundancy. If the length of the paraphrased sentence is much longer than the original sentence, it may indicate that additional information has been added. Conversely, if the length of the paraphrased text is significantly shorter than the original text, it may indicate that some information has been omitted or lost.

The design of these two features incorporates some prior knowledge as constraints. Specifically, if the paraphrased sentence and the source sentence are semantically similar, their text lengths should not differ significantly.

We obtain the SPScore-zh by fitting these eight features to the expert-annotated scores. As shown in Table 4 on the test set of $\text{CSPC}_{\text{core}}$, our evaluation metric exhibits a higher correlation compared to existing metrics.

Among them, we randomly selected 10% of the data from $\text{CSPC}_{\text{core}}$ as the test set and the rest is as train set. When training SPScore-zh, we only use the train set to ensure that there is no data leakage, thus ensuring the fairness of the comparative experimental results.

5 Paraphrase Generation

5.1 Establish Vocabulary

Regarding the Chinese sentence paraphrasing, we find that the traditional pre-training and fine-tuning paradigm performs poorly on such downstream tasks. This is primarily due to the specific characteristics of the sentence paraphrasing in Chinese. Firstly, sentence paraphrasing requires semantic consistency and diverse expressions between the original and paraphrased sentences. In the generation process of such pre-trained models, if we restrict the probability range of predicting the next word to achieve semantic consistency, the model is likely to preserve the original sentence without making any substantial changes. On the other hand, if we aim for diverse expressions, we need to expand the probability range of predicting the next

POS	Example	Words	HF Words	Ratio
General Noun	苹果	82823	17487	0.21
Verb	跑, 学习	34513	12750	0.37
Adjective	美丽	5918	4840	0.82
Idiom	百花齐放	8213	4419	0.54
Adverb	很, 非常	2226	2226	1.00
Person Name	杜甫	24316	1871	0.07
Geographical Name	北京	11862	1741	0.14
Other Noun-modifier	大型	2338	1068	0.45
Quantity	个, 粒	1035	1035	1.00
Other Proper Noun	诺贝尔奖	5065	1025	0.20
ALL	-	188655	53562	0.28

Table 5: Number of different POS in THUCNews and our vocabulary. HF words means high-frequency words.

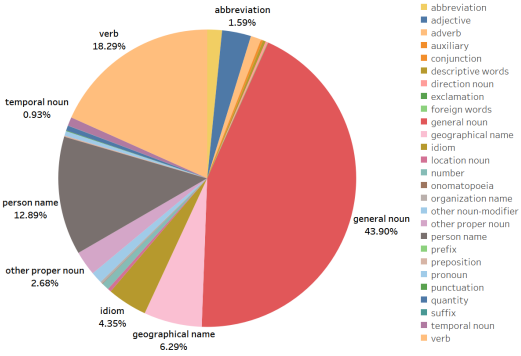

Figure 2: Distribution of different parts of speech in established vocabulary.

Example	Regular Expressions	Type
《红楼梦》	《(.*?)》	Book
3月15日	(\d1,10)\s*[\./月-]\s*(\d1,10)\s*日?	Date
(苹果是红的)	((.*?))	Explanation
“你吃饭了吗？”	“(.*?)”	Quote

Table 6: Fixed patterns and regular expressions. These phrases will be replaced by special tokens.

word, which can easily result in significant deviation from the original semantics.

This is mainly because the vocabulary of pre-trained models is usually fine-grained which takes many steps to generate synonyms, and this process often results in shifts during the generation process. However, sentence paraphrasing does not require such fine-grained generation but focuses more on synonymous replacements of entire words and changes in the position of the entire words.

To address this challenge, we develop a vocabulary based on high-frequency part-of-speech (POS) tagging. Based on the expert annotation results, Chinese sentence paraphrasing requires the extraction of key phrases that make up the sentence, including nouns, adjectives, adverbs, and other parts of speech. These key phrases need to be replaced with words of the same part of speech and similar semantics. Finally, these replaced words are combined with appropriate connecting words to form a complete sentence.

We first tokenize the THUCNews, a large-scale Chinese corpus, into words using LTP (Che et al., 2021), and classify the words based on part-of-speech tagging. For each part of speech, we add high-frequency words to our vocabulary as shown in Table 5. The final distribution of part-of-speech (POS) in the vocabulary is shown in Figure 2. It can be observed that there is a higher proportion of vocabulary selection of adjectives and verbs, which are related to the paraphrase. On the other hand, there is a lower proportion of vocabulary selection for unique nouns, etc.

There are two key points to paraphrase. First, the replaced words before and after should have the same part of speech. By ensuring a similar number of words of different parts of speech that form the sentence before and after replacement, the semantics of the keywords that form the sentence can be more preserved. This simplifies the Chinese sentence paraphrasing task, as it only requires using appropriate conjunctions to connect the replaced keywords into a sentence.

The second point is that the replaced words should be high-frequency words. In Chinese writing, high-frequency words are usually general words and are often selected as replacement words in Chinese paraphrasing. On the other hand, low-frequency words usually represent places, names, or other proper nouns, which are often retained in Chinese paraphrasing. Because there are hardly any alternative words for proper nouns, proper nouns are easily misinterpreted by language models, leading to changes in the original meaning when generating paraphrased sentences.

Therefore, selecting high-frequency words from each part of speech as the vocabulary for Chinese paraphrase models helps the language model understand the connection between words with similar meanings in the same part of speech. This allows the model to accurately identify replaceable words and their corresponding replacement words when generating paraphrased sentences.

5.2 Entity-aligned Tokenizer

A suitable vocabulary is not enough for a language model to be effective in paraphrasing Chinese sentences, an appropriate tokenization strategy is also required. As mentioned before, when it comes to paraphrasing Chinese sentences, not all key-

words are replaceable. This includes time, location, names, and proper nouns, etc.

To address this challenge, we design an entity-aligned tokenizer that performs entity alignment during tokenization. It extracts entities with fixed patterns as shown in Table 6, such as time, titles, and numbers, using regular expressions and incorporates corresponding special tokens into the vocabulary, such as a time token and a location token. These special phrases are replaced by special tokens during tokenization.

Additionally, since these embeddings of special phrases are shared across different sentences, the model does not need to learn their semantics during training but rather focuses on their positional information within the sentence. In this case, we can preserve these irreplaceable words entirely and achieve changes in sentence structure.

Furthermore, after entity alignment and tokenization, there may still be certain low-frequency words that are not present in the vocabulary and do not have fixed patterns. In this case, we divide them based on their length. For words with fewer than k characters, we split them into individual characters, while for words with k or more characters, we replace them with other special tokens. This is because, in terms of tokenization results, longer and infrequently occurring words are typically non-standard proper nouns, while shorter words often function as connectors or compounds, with their semantic information derived from the finer-grained characters that compose them. Therefore, we split them to enable the model to independently learn the meanings of these individual characters.

5.3 Three-stage Training

Through the specifically designed vocabulary and tokenization strategies mentioned above, we propose a strong baseline for Chinese sentence paraphrasing, training on our proposed $\text{CSPC}_{\text{auto}}$, and testing on $\text{CSPC}_{\text{core}}$ which is completely human-paraphrased and expert-annotated.

In addition, we adhere to two key points. First, the higher the quality of the data, the better training performance of the model. However, high-quality data is not always sufficient, and relying on such a small amount of high-quality data may result in a lack of generalization while incorporating low-quality data can interfere with the training process. Second, Chinese sentence paraphrasing is often performed incrementally, starting with small detailed changes and gradually making the paraphrase.

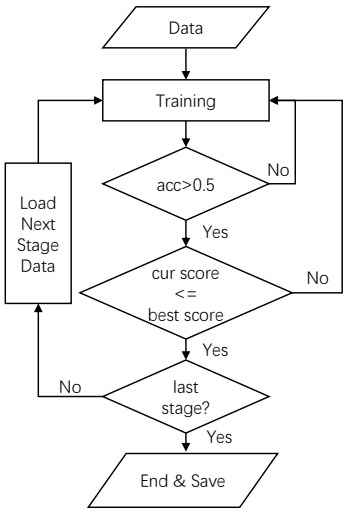

Figure 3: Three-stage training pipeline. Acc means the prediction accuracy of the language model. Cur score means the current SPScore-zh on the test set and the best score means the best SPScore-zh score that the model can achieve before.

Therefore, inspired by self-paced curriculum learning (Jiang et al., 2015), we propose a three-stage training strategy. We divide the dataset into three parts based on the scores and conduct three-stage training from low to high. Low-scored data typically has insufficient paraphrase or semantic variations, making them relatively easy samples for paraphrasing. High-scored data, on the other hand, involves more comprehensive changes while preserving high-quality semantics, making them more challenging to learn. Additionally, we embed the stage scheduling directly into the training process, enabling it to autonomously determine whether to progress to the next stage of learning without manual control as shown in Figure 3.

First, we filter lower-scoring sentence pairs as the first-stage data. During the training process, we ensure the readability of text generation by validating whether the accuracy of text prediction exceeds 0.5. We then establish an early stop strategy based on the SPScore-zh score. After the completion of the first-stage fitting, the training automatically proceeds to the next stage. In the second stage, we train on data with moderate scores from the dataset, and in the third stage, we train on data with higher scores. The training concludes after the completion of the third-stage fitting.

6 Case Study

The generation results can be seen in Table 7. It shows that our generation model can achieve ba-

	Original Sentence	Paraphrased Sentence	ST	SC	SR	SPScore-zh
Ours	相信爱读书的朋友或者居家办公的朋友都知道，一个舒适的工作和读书环境有多么的重要。那么这个环境是否舒适，就取决于我们书房的布置，尤其是写字台的布置。	相信喜欢读书的朋友们，或者是居家办公的朋友们都知道，一个人在工作和学习环境中有多舒服，对他来说有多么重要。那么这样的环境是否舒适，就完全取决于我们整个书房的布置了，更取决于我们工作室的布置了。	✓	✗	✓	0.646
	另外，写字台的布置也与风水息息相关，所以要分外谨慎。接下来就为大家揭秘哪些布局是有利于书房风水的吧。	另外，我们在办公室的布置也和风水有着密切的关系，因此对于办公空间来说，还是要格外小心一些。接下来就给大家揭晓一下，书房风水适合做哪些布局?	✓	✓	✓	0.931
	城市中的建筑大多是简易棚屋，用废弃的建筑材料做成，由于没有上下之分，一般都做成六面全有窗（也是门）的立方体，或者做成球形。	城市的建筑多为简单棚屋，用废弃的建材制成，由于没有上下之分，一般做成六面全是窗户（也是门）或呈球形的长方体。	✗	✗	✓	0.603
Human	走过去搂着汪珊说：“老婆，对不起，我错了，我不该这么说你。困难只是短暂的，再忍忍，孩子大一点就好了。”	走到汪珊身边，搂住她，轻声细语道：“宝贝，对不起，脱口而出的话真的太不恰当了。可是，困难也不过是短暂的，只要我们在一起陪伴孩子，他们长大就好。”	✓	✓	✓	0.735

Table 7: Case study of our paraphrase generation model.

Method	SPScore-zh↑	std.↓
Back-Translation	0.617	0.213
GPT-3.5 Turbo	0.641	0.218
Human	0.711	0.167
Ours	**0.724**	**0.157**

Table 8: Comparison between our paraphrase generation model and other paraphrase methods. It can be seen that our model can generate sentences with higher scores and more stable.

sic synonym replacement and sentence structure transformation, with some weakness in maintaining semantic consistency. This is mainly due to the presence of semantic errors in the back-translated corpus caused by the limitations of the translation model. It is necessary to enhance the discriminative ability of detailed semantics consistency in SPScore-zh. The SPScore-zh is only based on vector semantic similarity for comparison and filtering, making it difficult to detect such detailed errors. For example, the semantic similarity between a 球形长方体 and a 球形 is high.

In the first sentence of our generation samples, “环境是否舒适应取决于写字台的不止” while the paraphrased sentence changes it to “环境的舒适取决于工作室的布置” which alters the original semantics, thus not achieving semantic consistency. In the second example, changing “息息相关” to “密切关系” and modifying the original sentence structure with “因此” makes this example a high-scoring sentence paraphrase.

We also compared our generation model with other paraphrasing methods, as shown in Table 8. Our method achieves state-of-the-art results in terms of both generation quality and stability. Note that, as mentioned in Section 3.1, human paraphrasing often results in low-quality paraphrases due to work fatigue or limited language proficiency. Apart from that, human paraphrasing may also utilize other tools such as translation or generation models, as well as text errors that arise due to rewriting fatigue. In contrast, our generation model is not subject to these limitations, which is why it slightly outperforms human paraphrasing in overall scores.

7 Conclusion

In this paper, we emphasize the research gaps and various challenges in Chinese sentence paraphrasing, including the lack of automatic evaluation metrics which is aligned with human annotations, the absence of high-quality Chinese paraphrasing datasets, and the lack of language models capable of performing paraphrase generation.

To address these challenges, we propose SPScore-zh, a novel neural-based Chinese sentence paraphrasing evaluation metric that is highly aligned with human annotations. Additionally, we propose a large-scale and high-quality Chinese sentence paraphrasing dataset through automatic generation and filtering. Finally, using the established SPScore-zh and the high-quality dataset, we introduce a strong baseline for Chinese paraphrase generation that is capable of consistently producing high-quality paraphrased sentences. And we have already launched the paraphrase application programming interface (API) into production, which effectively validates the applicability of our method in practical applications.

8 Limitations

However, the evaluation metric SPScore-zh and the paraphrase generation model we propose are designed specifically for Chinese. In the future, we plan to achieve more language adaptation in paraphrase evaluation and generation. Additionally, our proposed baselines currently do not achieve significant results in structural transformation, and the degree of paraphrase generation is also not controllable, requiring further optimization.

References

Rahul Bhagat and Eduard H. Hovy. 2013. What is a paraphrase? *Comput. Linguistics*, 39(3):463–472.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: an open-source neural language technology platform for chinese. In *EMNLP (Demos)*, pages 42–49. Association for Computational Linguistics.

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. 2018. The BQ corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *EMNLP*, pages 4946–4951. Association for Computational Linguistics.

Yao Dou, Chao Jiang, and Wei Xu. 2022a. Improving large-scale paraphrase acquisition and generation. In *EMNLP*, pages 9301–9323. Association for Computational Linguistics.

Yao Dou, Chao Jiang, and Wei Xu. 2022b. Improving large-scale paraphrase acquisition and generation. In *EMNLP*, pages 9301–9323. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

James Ferguson, Colin Lockard, Daniel S. Weld, and Hannaneh Hajishirzi. 2018. Semi-supervised event extraction with paraphrase clusters. In *NAACL-HLT (2)*, pages 359–364. Association for Computational Linguistics.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *ACL (1)*, pages 6065–6075. Association for Computational Linguistics.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *EACL*, pages 1022–1033. Association for Computational Linguistics.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. 2015. Self-paced curriculum learning. In *AAAI*, pages 2694–2700. AAAI Press.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha P. Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Trans. Assoc. Comput. Linguistics*, 8:330–345.

Zibo Lin, Ziran Li, Ning Ding, Haitao Zheng, Ying Shen, Wei Wang, and Cong-Zhi Zhao. 2020. Integrating linguistic knowledge to sentence paraphrase generation. In *AAAI*, pages 8368–8375. AAAI Press.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *COLING*, pages 1191–1204. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka, and Eneko Agirre. 2022a. Principled paraphrase generation with parallel corpora. In *ACL (1)*, pages 1621–1638. Association for Computational Linguistics.

Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka, and Eneko Agirre. 2022b. Principled paraphrase generation with parallel corpora. In *ACL (1)*, pages 1621–1638. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *ACL (1)*, pages 866–876. Association for Computational Linguistics.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *EMNLP*, pages 3178–3190. Association for Computational Linguistics.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *ACL (2)*, pages 38–42. The Association for Computer Linguistics.

Moming Tang, Chengyu Wang, Jianing Wang, Cen Chen, Ming Gao, and Weining Qian. 2023. Parasum: Contrastive paraphrasing for low-resource extractive text summarization. In *KSEM (3)*, volume 14119 of *Lecture Notes in Computer Science*, pages 106–119. Springer.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *EMNLP (1)*, pages 90–121. Association for Computational Linguistics.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *NGT@EMNLP-IJCNLP*, pages 215–220. Association for Computational Linguistics.

Congle Zhang, Stephen Soderland, and Daniel S. Weld. 2015. Exploiting parallel news streams for unsupervised event extraction. *Trans. Assoc. Comput. Linguistics*, 3:117–129.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.