CPM: A large-scale generative Chinese Pre-trained language model

Wang Xiaoxing

Abstract

Pre-trined Language Models (PLMs) have proven to be beneficial for various downstream NLP tasks. Article we reproduce release the Chinese Pre-trained Language Model (CPM) with generative pre-training on large-scale Chinese training data.Howerver, CPM focus on generating general-purpose content, as they suffer from the risk of being toxic, overly complex, and unfair.Considering the above shortages, we take the first step to construct CCStories, a carefully selected&cleaned fair and high-quality dataset that contains 22K children's stories written in Chinese. We also implemented experiments to compare CPM and its fine-tuned variant on CCStories. The experimental results show that the fine-tuned model on CCStories generate more reasonable and safer stories that are more appropriate for children.

Keywords: Pre-trined Language Models, Story Generation, Children Story.

1 Introduction

Language and mental development are key aspects of early childhood education. Children's stories, through the characters and virtues shown in the content, provide moral models for kids to follow and help to develop their vocabulary by offering vivid illustrations[1]. To relieve the enormous workload of people writing, neural text generation has achieved remarkable success with the development of pretrained generation models[2-5], which are widely used in dialogue generation, question-answering systems, and story generation.

However, existing language generation models[4-6] focus on the generation of general-purpose content, yet there is a scarcity of high-quality research on generating children's stories for childhood education. Existing language generation models may perform well for the general domain, but they fall short in understanding and generating what is the most appropriate for children's educational purposes[7-8]. These works do not take into account the complexity and readability of the text generated by the model, etc. For example, they suffer from the risk of being toxic, overly complex, and unfair[**lemaignan2021unicef，charisi2021designing**]. The risk to children may result when building applications for children's education. An elaborate children's dataset is required to mitigate this risk

There are a number of corpora[4-5] that are collected online from Reddit[1] and articles on Wikipedia, etc.These datasets are usually from posts and comments of adult users, which contain complex words and sentences that children cannot understand, and even some toxic content. Language models learn positive features, but also negative ones like harmful and complex content that is not suitable for children.

[1] https://www.reddit.com/

Though there are few available children's textual datasets, they concentrate on content in English, shadowing the progress of story generation for children in other languages (e.g, Chinese.).In this situation, we take the first step to construct a fair, high-quality, and children-oriented Chinese story generation corpus,we take the first step to construct a fair, high-quality, and children-oriented Chinese story generation corpus,called CC-Stories, consisting of 22,200 children's stories in Chinese with titles. The dataset was gathered from public children's story websites. We ensure the quality of the corpus through automatic cleaning and manual cleaning. We also ensure that each story is complete and coherent without duplication by manually checking. We conduct experiments based on two widely used language models. We find that the problems of being toxic, overly complex, and unfair to children do exist in the stories generated by the general-purpose language models. With fine-tuning on CCStories dataset, the readability and safety of stories generated by fine-tuned models are improved significantly.

2 Related works

We discuss studies related to Chinese children's story generation, including (1) relevant story datasets, and (2) Chinese text generation models.

2.1 Story datasets.

ROCStories[9] proposed a dataset of 50K high-quality five-sentence stories. WritingPrompts[10] constructed a dataset containing 300K human-written stories paired with writing prompts. Both ROCStories and WritingPrompts have been widely used for story generation.There also are some Chinese story datasets. STORAL is composed of Chinese and English stories paired with morals[11]. LOT[12] constructs a Chinese story dataset with outlines for some tasks including Cloze Test, Sentence Position Prediction, Plot Completion, and Outline-conditioned Generation.

Unlike the above dataset, we provide a story dataset specifically for children that can facilitate the application of language models in the area of children's education.

2.2 Chinese text generation models

Text generation has achieved remarkable success with the development of pretrained generation models. Cui et al. propose a Chinese pre-trained language model MacBERT based on BERT[14], which mainly reduces the difference of the original BERT between training and fine-tuning. Based on BERT, Wei et al. propose NEZHA, which performs well in several Chinese language understanding downstream tasks, such as including named entity recognition, sentence matching, Chinese sentiment classification, etc. Sun et al. propose a new linguistic representation model enhanced by knowledge, called ERNIE, which exhibits strong knowledge inference on Chinese NLP tasks. The "Text-to-Text Transfer Converter" (T5)[6] performs well on various English NLP tasks utilizing a uniform text-to-text format and scale. Based on T5, Xue et al. propose mT5, which is pre-trained on a new dataset of Common Crawl with 101 languages including Chinese. Mengzi[18] is a multimodal Chinese pre-training model capable of performing a wide range of linguistic and visual tasks. Based on the transformer autoregressive model, Zhang et al. propose a large-scale Chinese pre-trained language model with

2.6 billion parameters and pre-trained on 100GB of Chinese data. Based on the transformer architecture, Zeng et al. propose Chinese pre-trained language models with up to 200 billion parameters, which are pre-trained on a 1.1 TB high-quality Chinese corpus.

3 Chinese PLM

CPM is a left-to-right Transformer decoder, which is similar to the model architecture of GPT . In order to adapt CPM to Chinese corpora, they build a new sub-word vocabulary and adjust the training batch size.

Vocabulary Construction: Previous works on Chinese pre-trained models usually adopt the sub-word vocabulary of BERT-Chinese (Devlin et al., 2019), which would split the input text to a character-level sequence. However, Chinese words usually contain several characters, and some important semantic meanings of words would be lost in the character-level sequence. To solve this problem, we construct a new sub-word vocabulary, containing both words and characters. For example, some common words would be added to the vocabulary.

Training Strategy: Since the sparseness of word distributions of Chinese is more serious than that of English, we adopt a large batch size to make the model training more stable. Compared to the batch size (1 million tokens) used in GPT-3 2.7B (Brown et al., 2020), our batch size (3 million tokens) is two times larger. For the largest model, which cannot be stored in a single GPU during training, we partition the model across GPUs along the width dimension to make the large-scale training available and reduce data-transfer among nodes.

4 Dataset

4.1 Dataset Collection

Our dataset is collected from several children's stories websites. The website links and the number of stories crawled are shown in Table 1.We collect all the articles and their related information on the websites, including titles, authors, reading suggestions, pronunciation of words (pinyin), etc. Articles include general stories, poems and science facts, etc. These articles are presented in different forms, some parts in text form, and some in pictures. Note that we only collect the text of the articles, which results in some of them being incomplete. The total number of candidate articles collected is 47,501. To ensure the quality of the data, two data cleaning processes are designed, including automatic cleaning and manual cleaning.

[2] http://www.tonghuaba.cn/

[3] http://www.quangushi.com/

[4] https://www.gushi365.com/

[5] https://www.etgushi.com/

[6] http://www.kuailegushi.com/

[7] http://www.xwjj.cn/

[8] https://www.qigushi.com/

[9] https://www.baobaogushi.com/

Website	**Num**
童话故事网[2] (Fairy Tale Network)	13,979
全故事网[3] ((Broad Story Network)	11,820
故事 365[4] (Story-365 Network)	6,268
儿童故事网[5] (Children’s Story Network)	6,206
快乐故事网[6] (Happy Story Network)	3,735
贝贝故事网[7] (Bebe Story Network)	2,569
七故事网[8] (Qi Story Network)	1,501
[9] (Baby Story Network)	1,423
Total	47,501

表 1: Sources of CCStories dataset including the websites, links, and the number of stories from each of the websites.

Categories	**Samples**
Fairy Tales	4842
Fables	4201
Folk Tales	2908
Puzzle Stories	1231
Mythology	391
Other Stories	8627

表 2: Category Statistics of CCStories.

4.2 Cleaning Process

To ensure the quality of the dataset, we design a series of rules to clean the data automatically. The steps are as follows: (1) Delete noisy characters: website name, links, pinyin, redundant space, newline, and special symbols (e.g., ♢, ■, -, |, 【, 】); (2) Replace uncommon punctuation marks: a) replace English apostrophes with Chinese apostrophes (. . . → .....) and b) replace continuous Chinese period with Chinese apostrophes (。。。。。。 → .....) c) Replace Chinese special quotation marks with Chinese inverted quotation marks (「」→ “”); (3) Removing content not relevant to the stories (mainly at the end/beginning of the story): story source, author information, reading suggestions, reading reflections, etc.

During the cleaning process, we found that there are redundant and irrelevant data in many articles, which are difficult to remove through automatic data cleaning processing. It is also difficult to filter out non-story content (e.g., poetry) and incomplete content by automatic cleaning. Thus, we recruit eight volunteers as annotators to manually clean the data.

4.3 Dataset Statistics

CCStories contains many categories of stories, such as fairy tales, fables, mythology, etc. The statistics of the story categories are shown in Table 2. Other stories mainly include bedtime stories, educational stories, and philosophical stories. We also categorize the stories by the number of words, the statistic is shown in Table 3.

Length	0-500	500-1000	1000-2000	>2000
Samples	7186	7880	4141	2993

表 3: Number of samples with different lengths.

Datasets	Samples	#Char	#Word	#Sent
Train	20000	1062.59	654.73	44.87
Validation	1200	1052.93	648.28	43.70
Test	1000	1024.36	632.38	43.99

表 4: CCStories Statistics. #Char/ #Word/ #Sent denote the average numbers of characters/words/sentences, respectively.

We split the dataset into a training set (20,000 samples), a validation set (1,200 samples), and a test set (1,000 samples) in the proportion of 90%, 5.5%, and 4.5%, respectively. To better understand the distribution of the dataset, we calculate the average number of characters, words, and sentences. In addition, to count the number of words, we first split the sentences by Jieba[10] for word tokenization and then count the average number of words. Table 4 shows the statistics of CCStories.

5 Implementation details

5.1 Comparing with released source codes

We use the published fine-tuning code. We have added code related to the experimental evaluation assessment. In addition, We construct a fair, high-quality, and children-oriented Chinese story generation corpus,called CC-Stories, consisting of 22,200 children’s stories in Chinese with titles. We use the datasets we provide to train and test.

5.2 Experimental environment setup

Due to computational resource constraints, we use the version with 2.6B parameters of CPM as baselines. We evaluate the following model: CPM, CPM-FT (CPM fine-tuned on CCStories).

For fine-tuning settings, we train model on CCStories up to 10 epochs. We set the batch size to 2 for CPM . We use the Adam optimizer with an initial learning rate of $1.5e^{-4}$. For inference, we apply top-p=0.85 and the temperature of 1.0 to generate stories. The maximum length of the generated text is 300.All results are based on one NVIDIA A100 GPU. It cost less than 48 hours for fine-tuning each model on CCStories.

5.3 Case study

The model input requires a title and a prompt, and then the model finishes a story based on the prompt. We show some cases generated by the models in Table 5. These examples show that the original pretrained baseline model is capable of generating stories. However, the generated stories are not children’s stories and their content contains some violent and biased words (e.g., 杀人 (murder).) that are not suitable for children to read. This may be attributed to the fact that the pretrained corpus has negative and sensitive content included. The stories generated by the models fine-tuned on CCStories are simpler, more understandable, and with fewer negative words.

5.4 Main contributions

We propose a new Chinese Children Story dataset containing 22,200 stories called CCStories. We conduct a series of experiments to show that general language generation models fine-tuned on our dataset produce

[10] https://github.com/fxsjy/jieba

Prompt: 有一天, 这里来了一个黑人医生。(Translation: One day, a black doctor came here.)

CPM

有一天, 这里来了一个黑人医生。他到了这里, 就被当作人质, 关押在牢里, 用刀片割下了他的喉咙, 然后把他扔进医院的废墟。再有几天, 所有的人都开始对那位黑人医生怀恨在心。直到四年之后, 记者才知道, 那位黑人医生为什么不出来 杀人 了。那个黑人医生是 黑鬼 的医生, 为了报仇, 他把自己的脸贴在一个白色的鼻子上, 然后被扔进了自己的屋子里。. . .

(Translation: One day, a black doctor came here. When he arrived, he was taken hostage, imprisoned in a cell, his throat was cut with a blade, and he was thrown into the ruins of the hospital. For a few more days, all the people began to hold a grudge against the black doctor. It was not until four years later that the reporter learned why the black doctor had not come out to kill. The black doctor was a nigger's doctor, and to avenge his death, he put his face on a white nose and was thrown into his house. . . .

CPM-FT

有一天, 这里来了一个黑人医生。他的名字叫巴尔特克。他是个医生, 但是他的医术谁也不知道, 所以他最好还是自己种植药材。他听人说, 有一种草药, 在开罗城 • "药" 里有一种, 样子很怪, 像长着癞蛤蟆的脸。这种药叫 "麻松"。巴尔特克给这个名叫 "麻松" 的人送了一麻袋 "麻松", 请他到啤酒店去, 对他的医术有很深造诣。. . . (Translation: One day, a black doctor came here. His name was Baltke. He was a doctor, but his medical skills were unknown to anyone, so it was best for him to grow his own herbs. He heard that there was a kind of herb that was found in the "medicinal spring" in Cairo, which looked strange, like the face of a toad. This medicine is called "Masson". Baltke sent a sack of "Masson" to the man named "Masson" and asked him to go to the beer store, which had a deep knowledge of his medical skills. . . .

表 5: Examples of generated stories using language generation models based on the same prompt. This is part of the content generated by the models, the . . . indicate the truncated content. It can be observed that the stories generated by the original models, i.e., CPM and PanGu, are more unfair, e.g., being longer, and even toxic for children, e.g., containing inappropriate works such as raping and killing. While stories generated by models finetuned on CCStories are more suitable for children.

Models	**Readability**	
	Character	Token
CPM	0.432	0.605
CPM-FT	**0.516**	**0.708**

表 6: The results of children's unfairness in terms of readability.

content that is more suitable and safer for children to read.

6 Results and analysis

For readability, we calculates the ratio of words at different vocabulary levels of the text output by the model. Low-level terms are assigned large weights, while high-level terms are assigned low weights. So higher scores indicate higher readability. The readability scores are shown in Table 6. Experimental results show that the readability of fine-tuned models is higher than original models at both the character and token levels.

We use the Chinese Offensive Language Detection (COLD)[11] to classify whether the text generated by the model is toxic or not. Specifically, we use two different datasets (offensive and non-offensive) as inputs to test the toxic score of the model separately. The offensive input is more likely to induce the model to generate toxic content than the non-offensive input. The offensive dataset comes from the training set of COLD, and we extract 2,090 sentences with less than 20 words to construct the dataset. The non-offensive dataset comes from the test set of CCStories, and we construct the non-offensive dataset by taking the title and the first 15 words of each story. We take the first 120 words of the output for toxic classification. We report the toxicity

[11] https://github.com/thu-coai/COLDataset

Models	**Toxicity**	
	Non-Offen	Offen
CPM	0.122	0.410
CPM-FT	**0.099**	**0.143**

表 7: The results of children's unfairness in terms of Toxicity. **Bold** indicates the least toxicity.

of a model as the proportion of its generated toxic content to all its content. Therefore, a larger toxicity score indicates that the model is more toxic.The results are shown in Table 7.

Table 8 shows the story generation performance of different models. BLEU-1 (B-1), BLEU-2 (B-2), and Coverage (Cov) measure the n-gram similarity between the text generated by the model and the ground truth label text. BLEURT (BR) and COMET-22 (COM), are used to measure the sentence-level similarity between the text generated by the model and the ground truth label text. Distinct-4 (D-4) and Repetition-4 (R-4) measure the diversity of the content generated by the model, with larger D-4 values indicating more diverse text, and smaller R-4 values indicating a more diverse text. We can see from Table 8, BLEU-1, BLEU-2, Distinct, Repetition, and Coverage scores are improved by fine-tuning on CCStories. Only the Distinct score of PanGu dropped from 99.79 to 96.86 after fine-tuning, but this drop is acceptable. The results from the BR metric show that after fine-tuning with our data, the performance of CPM has improved, but the performance of PanGu has decreased. Therefore, the results show that fine-tuned models are more likely to generate children's stories than original models.

Models	**B-1**	**B-2**	**D-4**	**R-4**	**Cov**	**BR**	**COM**
CPM	10.97	2.04	78.32	84.00	27.21	35.08	-118.07
CPM-FT	**22.09**	**8.20**	**86.95**	**79.80**	**38.70**	**36.94**	**-39.70**

表 8: The story generation performance of models. **Bold** indicates the best performances.

7 Conclusion and future work

We proposed a new Chinese Children' s Stories corpus CCStories. which contains at least five types of stories with a total of 22,200 stories. We ensured the quality of the dataset through automatic cleaning and manual cleaning. We also implemented experiments to compare Chinese pretrained language CPM and its fine-tuned variant on CCStories. The experimental results show that the general language generation model tends to generate content that is hard for children to understand and toxic. CCStories helps the model generate safer stories that are more appropriate for children, thus mitigating the potential unfairness to children. In the future, we want to investigate more language generation models especially models with larger numbers of parameters with our CCStories dataset.

References

[1] LIU R, NIKOLIC P K. Mutltimodal AI Companion for Interactive Fairytale Co-creation[J]. arXiv preprint arXiv:2112.00331, 2021.

[2] PETERS M E, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations[C/OL]

//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018: 2227-2237. https://aclanthology.org/N18-1202. DOI: 10.18653/v1/N18-1202.

[3] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.

[4] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[5] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[6] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer.[J]. J. Mach. Learn. Res., 2020, 21(140): 1-67.

[7] LUCKIN R, HOLMES W, GRIFFITHS M, et al. Intelligence unleashed: An argument for AI in education[J]., 2016.

[8] LA FORS K. Toward children-centric AI: a case for a growth model in children-AI interactions[J]. AI & SOCIETY, 2022: 1-13.

[9] MOSTAFAZADEH N, CHAMBERS N, HE X, et al. A corpus and cloze evaluation for deeper understanding of commonsense stories[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 839-849.

[10] FAN A, LEWIS M, DAUPHIN Y. Hierarchical neural story generation[J]. arXiv preprint arXiv:1805.04833, 2018.

[11] GUAN J, LIU Z, HUANG M. A Corpus for Understanding and Generating Moral Stories[J]. arXiv preprint arXiv:2204.09438, 2022.

[12] GUAN J, FENG Z, CHEN Y, et al. LOT: A Story-Centric Benchmark for Evaluating Chinese Long Text Understanding and Generation[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 434-451.

[13] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[14] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[15] WEI J, REN X, LI X, et al. Nezha: Neural contextualized representation for chinese language understanding[J]. arXiv preprint arXiv:1909.00204, 2019.

[16] SUN Y, WANG S, LI Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.

[17] XUE L, CONSTANT N, ROBERTS A, et al. mT5: A massively multilingual pre-trained text-to-text transformer[J]. arXiv preprint arXiv:2010.11934, 2020.

[18] ZHANG Z, ZHANG H, CHEN K, et al. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese[J]. arXiv preprint arXiv:2110.06696, 2021.

[19] ZHANG Z, HAN X, ZHOU H, et al. CPM: A large-scale generative Chinese pre-trained language model [J]. AI Open, 2021, 2: 93-99.

[20] ZENG W, REN X, SU T, et al. PanGu-$\alpha $: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation[J]. arXiv preprint arXiv:2104.12369, 2021.