AI Open 2 (2021) 65–68

Contents lists available at ScienceDirect

AI Open

journal homepage: www.sciencedirect.com/journal/ai-open

WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models

Sha Yuan[a], Hanyu Zhao[a], Zhengxiao Du[b], Ming Ding[b], Xiao Liu[b], Yukuo Cen[b], Xu Zou[b], Zhilin Yang[c], Jie Tang[b,*]

[a] *Beijing Academy of Artificial Intelligence, China*
[b] *Department of Computer Science and Technology, Tsinghua University, China*
[c] *Recurrent AI, China*

ARTICLE INFO

Keywords:
Pre-trained language models
Chinese corpus
Transformer-XL

ABSTRACT

Using large-scale training data to build a pre-trained language model (PLM) with a larger volume of parameters can significantly improve downstream tasks. For example, OpenAI trained the GPT3 model with 175 billion parameters on 570 GB English training data, enabling downstream applications building with only a small number of samples. However, there is a lack of Chinese corpus to support large-scale PLMs. This paper introduces a super large-scale Chinese corpora WuDaoCorpora, containing about 3 TB training data and 1.08 trillion Chinese characters. We also release the base version of WuDaoCorpora, containing about 200 GB training data and 72 billion Chinese characters. As a baseline, we train a model transformer-XL with 3 billion parameters on the base version to test the corpora's effect. The results show that the models trained on this corpora can achieve excellent performance in Chinese. The data and model are available at https://data.wudaoai.cn and https://github.com/THUDM/Chinese-Transformer-XL, respectively.

1. Introduction

A large number of studies have consistently shown that training pre-trained language models (PLMs) on large corpora can help it learn general language representation (Devlin et al., 2019; Zhang et al., 2012; Liu et al., 1907; Lan et al., 1909; Yao et al., 1909; Sun et al., 1907; Yang et al., 2019). Fine-tuning based on PLMs can improve downstream tasks' effects and avoid training models of downstream tasks from scratch (Qiu et al., 2003). The development of computing power and continuous innovation of sparse training methods support the growth of model scale. PLMs with more parameters inevitably require more data to support the training process. For example, OpenAI uses 570 GB data to train the GPT3 with 175 billion parameters (Brown et al., 2005); Google uses 750 GB colossal clean crawled corpus (C4) to train the 1.6-trillion--parameter Switching Transformer (Fedus et al., 2021).

With the rapid development of natural language processing (NLP) technologies, constructing large corpora becomes increasingly important. The quality of NLP models strongly relies on the scale of corpus used for model training (He et al., 2018; Duan et al., 1912; Cui et al., 2019; Qi et al., 2001). For example, EleutherAI built a 1.2 TB English corpus including material like Books3, PubMed Central, and FreeLaw. Many models achieve better performance in NLP tasks by using this corpus as training data. However, the largest Chinese corpus previously is the CLUECorpus2020 (Xu et al., 2003) with only 100 GB data, which suggests it is difficult to satisfy the demand of training high-quality model for Chinese NLP tasks. In order to solve this problem, we constructed a 3 TB Corpora in Chinese named WuDaoCorpora, which is mainly composed of web data. Moreover, models trained by our dataset can have better generalization ability because the Corpora contains various data types including news, post bar comments, encyclopedia information, etc.

More specifically, WuDaoCorpora contains a 3 TB Chinese corpus collected from 822 million Web pages. These data sources have been structured in a unified format. Each web page's text content can be obtained through the 'content' index, which guarantees the direct use of corpus for model pre-training without extra operations. In comparison with other corpora, WuDaoCorpora performs better in personal information protection. The models' outputs may expose private information in the training data. To prevent the models trained with WuDaoCorpora from privacy disclosure, we delete all personal data in WuDaoCorpora.

[*] Corresponding author.
E-mail address: jietang@tsinghua.edu.cn (J. Tang).

https://doi.org/10.1016/j.aiopen.2021.06.001
Received 9 March 2021; Accepted 4 June 2021
Available online 15 June 2021

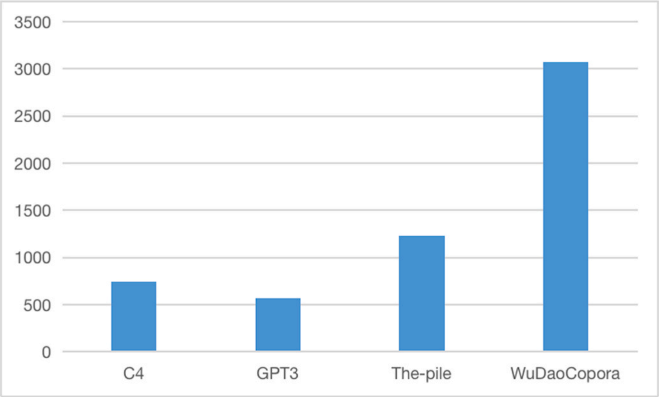

Fig. 1. Comparison with other Chinese corpora.

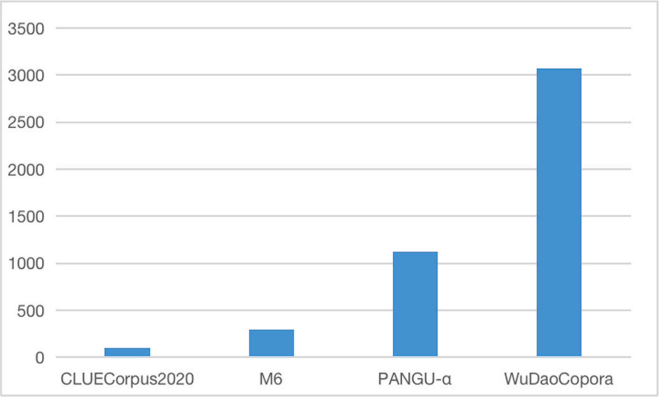

Fig. 2. Comparison with other language corpora.

Moreover, there is an index label in every piece of web page data in WuDaoCorpora to annotate which field it belongs. Finally, a three billion-parameter Chinese PLM named transformer-XL is trained based on the WuDaoCorpora base version library.

To summarize, this paper makes the following contributions:

- We construct the world's largest Chinese Corpora WuDaoCorpora, which contains 3 TB of Chinese data. It benefits model pre-training, word embedding, etc.
- In the construction of WuDaoCorpora, we focused on removing personal privacy information in data, which significantly reduces the risk of personal privacy leakage.
- The largest Chinese PLM transformer-XL is open-source, and its few-shot learning ability has been demonstrated.

2. Relation work

Corpora are essential resources in NLP tasks. Early released corpora for PLMs are in English. For example, Zhu et al. proposed a Toronto Books Corpus (Zhu et al., 2015), which extracts the text from eBooks with the size of 4 GB. Later, for the convenience of collating process and the variety of the data source, some corpus datasets are based on Wikipedia. Wikipedia dataset[1] contains 16 GB of text data extracted from encyclopedia articles. But Wikipedia is only one kind of article format and can not represent the whole language environment. To get more diverse text, many researchers use web-based text to construct the corpus. Radford (Radford et al., Sutskever) proposed a 17 GB WebText-like English corpus using the content aggregation website Reddit. Raffel proposed the C4 dataset (Raffel et al., 2020), a 750 GB English dataset consists of hundreds of gigabytes of clean English text scraped from the web.

At present, there are some Chinese corpora for PLMs. THUCTC[2] is a Chinese text classification toolkit accompanying by a 2.19 GB dataset containing 740,000 news documents. Li et al. (Li, 2003) pre-trained the GPT model on two Chinese corpora, Wikipedia2[3] (1.7B words) and Chinese News (9.2B words), and used a small dialogue dataset of Chinese to fine-tune the speech data to generate dialogue. Wang et al. (2008) established a clean Chinese dialogue dataset LCCC. This dataset has two versions, LCCC-base and LCCC-large. LCCC-base is filtered based on 79 million conversations crawled from Weibo, while LCCC-large is filtered from a combination of Weibo data and other Chinese corpora. It is worth mentioning that the dataset uses a set of rules and a classifier trained on 110 thousand manually labeled dialogue pairs to ensure the quality of the dataset. Compared to the English corpus, the Chinese dataset for PLMs is quite limited from the dataset size to the variety.

Table 1
Example of WuDaoCorpora.

Example

{
"id": "200023",
"url":"http://news.cri.cn/20180609/6ba0cfde-e78e-8577-b98f-200292-2b7a26.html",
"dataType": news,
"title": "3D微视频-共同家园 (3D Micro Video - Common Homeland)",
"content":" 黄海之滨，青青岛城。八方宾客齐聚，上合组织青岛峰会今天举行。人民日报推出3D微视频，共同家园。让我们跟随习近平主席的同期声，以全新方式讲述上合故事。诞生17年，上海合作组织日益壮大。从地区安全到经贸合作，从人文交流到开放包容，上合组织合作硕果累累。千帆过尽，不忘初心。上海精神始终贯穿上合组织发展历程。携手建设共同家园，中国智慧为上合组织注入新的发展动力。新的发展方位，新的历史起点，上合组织发展迎来历史性机遇。从青岛再度扬帆出发，上合组织未来发展新蓝图正在绘就。青青之岛，和合共生。(On the shore of the Yellow Sea, Green Island City. The Shanghai Cooperation Organization (SCO) Qingdao Summit opens today. People's Daily launches 3D micro video, Common Home. Let's follow the voice of President Xi Jinping and tell the story of Shanghai Ho in a new way. Since its founding 17 years ago, the Shanghai Cooperation Organization has been going from strength to strength. From regional security to economic and trade cooperation, from people-to-people exchanges to openness and inclusiveness, the SCO cooperation has yielded fruitful results. After a thousand sails have been completed, we will never forget our original aspiration. The Shanghai Spirit has been throughout the development of the SCO. Joining hands to build a common home, the Chinese wisdom has injected new impetus into the development of the SCO. A new development juncture and a new historical starting point represent a historic opportunity for the SCO's development. Starting from Qingdao, we are drawing up a new blueprint for the SCO's future development. Green island, harmonious coexistence.)"
"dataCleanTime": "2021-01-09 00:59:15"
}

3. Dataset description

Before the release of WuDaoCorpora, no TB-level Chinese corpus can be used directly for PLM training. Fig. 1 shows the comparison of scale between WuDaoCorpora and other Chinese corpora before it. When comparing worldwide, our dataset also has a larger scale than any other corpora currently (showed in Fig. 2). We use 3 billion raw web data to construct WuDaoCorpora. After a series of filtration and processing, we finally obtain 3 TB data in a unified format. Table 1 gives an example of a piece of data. In the next section, we will introduce the process of corpus construction in detail.

4. Dataset construction

A large-scale and high-quality corpus is the basis of self-supervised learning. To fill in the blank of TB-level Chinese corpus, we use 3

[1] https://tensorflow.google.cn/datasets/catalog/wikipedia.
[2] https://github.com/thunlp/THUCTC.
[3] https://github.com/brightmart/nlpChinese corpus/.

billion web pages as the original data source and extract text content from those with high text density. The extracted text contains many extra characters that damage the content's integrity and smoothness, such as the web page identifier, abnormal symbols, and garbled code. Additionally, sensitive information and personal privacy information exist in the text content extracted from some web pages, which may cause adverse tendencies and information leakage problems in trained models. To solve these problems, we developed a set of processing procedures in the data cleaning process to improve the corpus' quality.

- We evaluate every data source's quality before text extraction and ignore those web pages whose text density is lower than 70%.
- Because the phenomenon of text reposting is common on web pages, we use the simhash algorithm to remove those duplicated contents.
- Web pages with few words usually suggest that they do not contain meaningful sentences. These web pages are not suitable for training language models. If a web page contains less than 10 Chinese characters, we ignore it.
- Sensitive information like dirty words, seditious comments, and other illegal contents has adverse effects on building a harmonious and positive social environment. We exclude web pages that contain the above contents.
- To protect everyone's privacy security to the greatest extent, we use Regular Expression to match private information (i.e., identity number, phone number, qq number, email address, etc.) and remove them from the dataset.
- Incomplete sentences can be problematic in model training. We use punctuation marks (i.e., period, exclamation mark, question mark, ellipsis) to divide extracted texts and delete the last segment, which can be incomplete sometimes.
- Because of the breach of the W3C standard of some web pages, text extracted from them can be garbled. To exclude garbled contents in our corpus, we filter web pages with high-frequency garbled words and use a decoding test for double-checking.
- Since there are Chinese characters in both the simplified version and the traditional version, we transform those traditional characters to simplified versions to make the character format in our corpus unified.
- To guarantee the smoothness of extracted text, we remove those abnormal symbols (i.e., Emoji, logo, etc.) from web pages.
- To avoid the existence of long-length non-Chinese contents in our dataset, we exclude those web pages that contain fragments with more than ten successive non-Chinese characters.
- Since web page identifiers like HTML, Cascading Style Sheets (CSS), and Javascript are unhelpful for language model training, we remove them from extracted texts.
- Because using a space to separate two Chinese characters is not needed, we delete all spaces in every sentence to normalize our corpus.

5. Model

The structure of the model used in this paper is similar to the GPT2 model, which adopts the decoder of language model Transformer. The approach for predicting words in the GPT2 model is to compute the maximum likelihood of the conditional probability of present words given all history information. Specifically, for a series of token $U = \{u_1, u_2, \ldots, u_n\}$, the object function of the language model is: $L(U) = \sum_{i=1} \log P(u_i|u_{i-k}, \ldots, u_{i-1,\theta})$, where k is the window size and θ is the conditional parameter. Therefore, the $(n+1)^{th}$ token can be generated by considering all previous information.

Our model is promoted based on GPT2 in the following three aspects. 1) The language model Transformer is substituted with Transformer-XL to process longer series of tokens (Dai et al., 2019). 2) We reduced our model's window size to 512 and increased 256 unit length memories, which guarantees that the model can process longer token sequence than the previous model with window size of 1024. 3) We use Zero Optimizer in DeepSpeed to train a large model without applying the method of parallel training. According to the above promotions, we train a PLM with 3 billion parameters, being the largest Chinese language PLM presently.

Table 2
Example of transformer-XL.

Context	冬天，中国哪座城市最适合避寒？问题描述：能推荐一些国内适合冬天避寒的城市吗？回答用户：我的名字甜不甜 Which Chinese city is the best place to escape the cold in winter? Question description: Can you recommend some domestic cities suitable for winter cold? User: My Name Sweet.
transformer-XL	当然是敦煌啦! 我的家乡,一座离天堂最近的天府之国——以敦煌为中心,四周汇聚着雄伟壮观的中国石窟艺术宝库 最富盛名的莫高窟是断代的开凿,也因此成为未被盗掘过的天堂文物,与闻名世界的敦煌莫高窟隔着1620公里与沙漠隔着200公里的距离,这数字意味着敦煌的景象总是举世无双,中国数得清的历史岩石画,历尽沧桑的宗教绘画,艺术瑰宝,丝绸连天,丝路长河,"铁马秋风塞北,杏花春雨江南"的人物画像,只有身临其境才能更好地感受到的,最为真实的色彩和韵味。Dunhuang, of course! My hometown, a land of abundance closest to heaven, is centered in Dunhuang, surrounded by the magnificent treasure house of Chinese grotto art. The most famous Mogao Grottoes is the excavation of the period, so it has not been excavated by the heaven cultural relics, and the world-famous Dunhuang Mogao Grottoes are separated by 1620 km and 200 km from the desert, this figure means that the scene of Dunhuang is always unparalleled.

Table 3
Transformer-XL poetry examples.

Context	次韵姜山老游南湖 The old man Jiang Shan looked around in the South Lake 作者：宋 蔡襄 Poet: Caixiang Song
transformer-XL	梦里清颍西遇春，浪飞星斗似飞尘。 世间万事略无定，南湖桃李却分明。 I dreamed of the clear Ying water flowing west to meet the spring breeze. The waves were rolling and the stars in the sky like flying dust. Many things in this world have no definite target, only the peach blossom and plum blossom in the South Lake are relatively open and can be seen very clearly.

The layers of all our models are set to 24, and the number of heads is set to 16. The dimension of word embedding in our model are 1024. As for training details, the warmup epoch and batch size are 0.01 and 8, respectively, and the maximum learning rate is set to be 1.5e-5. Besides, the gradient clip is set to 1. All models are trained with a vocabulary of 13,088 Chinese characters on 64 NVIDIA Tesla V100 GPUs.

6. Experiments

In this section, we will test the performance of transformer-XL on language modeling tasks, such as question answering. First, we randomly sample around 2000 pieces of text data from Zhihu[4] and Baidu Baike[5] and use them as a test set. Then, we apply evaluation metrics to rate the model output.

The measure used for evaluation here is per-word perplexity (ppl), which evaluates the cognition degree. A lower ppl value means less confusion and therefore represents a high-level cognitive ability of a model. After the experiment, the ppl of GPT2 is 7.24, while our model achieve a significantly better score at 4.52.

Unlike traditional language modeling tasks, such as predicting a single word of interest, completing a sense or paragraph, we hope to explore the cognitive ability of transformer-XL, so we designed two

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

[5] https://baike.baidu.com/.

tasks: character setting dialogue and poetry writing.

Two applications are designed based on our language model. The first one is character setting dialogue. In the application of character setting dialogue, users input a question, including a description of the question and a designated responder. The model can generate answers that seem to accord with the initial setting of the responder's status. Table 2 gives one examples of character setting dialogue.

The second application is poem writing. Title of poem and the designated author are used as model inputs and transformer-XL will automatically generate relevant poems. Table 3 gives one examples of model-writing poetry.

7. Conclusion

In this paper, we introduce a super large-scale Chinese corpus for PLMs, WuDaoCorpora. It is known to be the largest Chinese corpus so far, which can be used directly to train PLMs. We filter the data source by filtering web pages according to DOM tree integrity, and text proportion characteristics during the data set construction. Also, a complete data set cleaning pipeline is developed, which mainly focuses on the clearance of privacy data to avoid personal privacy disclosure at the data level. Besides, every piece of data in WuDaoCorpora is annotated with a field tag, making it convenient to extract data in a specific field from the dataset to research the corresponding area.

To improve the cognitive ability of the PLMs, we will build the next version of WuDaoCorpora in the future. The corpus will contain more dialogue corpus and common sense knowledge, which will lay a solid data foundation for improving the cognitive ability of PLMs.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61806111, NSFC for Distinguished Young Scholar under Grant No. 61825602 and National Key R&D Program of China under Grant No. 2020AAA010520002.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krüger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2005. Language models are few-shot learners. ArXiv abs, 14165.

Cui, Y., Liu, T., Xiao, L., Chen, Z., Ma, W., Che, W., Wang, S., Hu, G., 2019. In: A Span-Extraction Dataset for Chinese Machine Reading Comprehension. EMNLP-IJCNLP.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R., 2019. In: Transformer-xl: Attentive Language Models beyond a Fixed-Length Context. ACL.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT.

Duan, X., Wang, B., Wang, Z., Ma, W., Cui, Y., Wu, D., Wang, S., Liu, T., Huo, T., Hu, Z., Wang, H., Liu, Z., 1912. Cjrc: a reliable human-annotated benchmark dataset for Chinese judicial reading comprehension. ArXiv abs, 09156.

Fedus, W., Zoph, B., Shazeer, N., 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity.

He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., Liu, X., Wu, T., Wang, H., 2018. Dureader: a Chinese machine reading comprehension dataset from real-world applications. In: QA@ACL.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 1909. Albert: a lite bert for self-supervised learning of language representations. ArXiv abs, 11942.

Li, P., 2003. An empirical investigation of pre-trained transformer language models for open-domain dialogue generation. ArXiv abs, 04195.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 1907. Roberta: a robustly optimized bert pretraining approach. ArXiv abs, 11692.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A., 2001. Imagebert: cross-modal pre-training with large-scale weak-supervised image-text data. ArXiv abs, 07966.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X., 2003. Pre-trained models for natural language processing: a survey. ArXiv abs, 08271.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models Are Unsupervised Multitask Learners 24.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv, 10683, 1910.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., Wang, H., 1907. Ernie 2.0: a continual pre-training framework for language understanding. ArXiv abs, 12412.

Wang, Y., Ke, P., Zheng, Y., Huang, K., Jiang, Y., Zhu, X., Huang, M., 2008. A large-scale Chinese short-text conversation dataset. ArXiv abs, 03946.

Xu, L., Zhang, X., Dong, Q., 2003. Cluecorpus2020: a large-scale Chinese corpus for pre-training language model. ArXiv abs, 01355.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V., 2019. Xlnet: generalized autoregressive pretraining for language understanding. In: NeurIPS.

Yao, L., Mao, C., Luo, Y., 1909. Kg-bert: bert for knowledge graph completion. ArXiv abs, 03193.

Zhang, Z., Han, X., Zhou, H., Ke, P., Gu, Y., Ye, D., Qin, Y., Su, Y.-S., Ji, H., Guan, J., Qi, F., Wang, X., Zheng, Y., Zeng, G., Cao, H., Chen, S., Li, D., Sun, Z., Liu, Z., Huang, M., Han, W., Tang, J., Li, J.-Z., Zhu, X., Sun, M., 2012. Cpm: a large-scale generative Chinese pre-trained language model. ArXiv abs, 00413.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. ICCV.