Chinese WPLC: A Chinese Dataset for Evaluating Pretrained Language Models on Word Prediction Given Long-Range Context

Huibin Ge[†], **Chenxi Sun**[†], **Deyi Xiong**[†], and **Qun Liu**[§]
[†] College of Intelligence and Computing, Tianjin University, Tianjin, China
[§] Huawei Noah's Ark Lab, Hong Kong, China
{gehuibin,cxsun,dyxiong}@tju.edu.cn
qun.liu@huawei.com

Abstract

This paper presents a Chinese dataset for evaluating pretrained language models on **W**ord **P**rediction given **L**ong-term **C**ontext (Chinese *WPLC*). We propose both automatic and manual selection strategies tailored to Chinese to guarantee that target words in passages collected from over 69K novels can only be predicted with long-term context beyond the scope of sentences containing the target words. Dataset analysis reveals that the types of target words range from common nouns to Chinese 4-character idioms. We also observe that linguistic relations between target words and long-range context exhibit diversity, including lexical match, synonym, summary and reasoning. Experiment results show that the Chinese pretrained language model PanGu-α (Zeng et al., 2021) is 45 points behind human in terms of top-1 word prediction accuracy, indicating that Chinese WPLC is a challenging dataset. The dataset is publicly available at https://git.openi.org.cn/PCL-Platform.Intelligence/Chinese_WPLC.

1 Introduction

Predicting a target word from previous context, especially long-range context, is a long-standing challenging problem in natural language processing. A variety of large-scale datasets such as CNN/Daily Mail (Hermann et al., 2015), Who-did-What (Onishi et al., 2016) and CMRC-2017 (Cui et al., 2018) have been developed to examine the capability of machines in word prediction. However, the majority of such datasets have not undergone a thorough manual testing whether a target word can only be predicted from long-range dependencies except for LAMBADA (Paperno et al., 2016). This dataset provides a benchmark testbed where a target word can be easily predicted with long-range context but cannot with only context words in the sentence where the target word is located.

Partially inspired by LAMBADA, we create Chinese WPLC, a dataset for evaluating powerful pretrained language models on word prediction with long-range context. The passages used in our dataset are carefully extracted from over 69K Chinese novels following a procedure mixed with automatic and manual selection. Significant differences from LAMBDA lie not only in language (English vs. Chinese), but also in the following two aspects:

- LAMBADA filters out relatively easy passages with weak language models, e.g., RNN, 4-gram and feed-forward neural language models, which makes it an outdated dataset for current state-of-the-art pretrained language models as target words in many left passages may be easily predicted by large-scale pretrained models. Additionally, the original raw data used by LAMBADA may potentially appear in the training set of current pretrained models (Brown et al., 2020). To tackle the aforementioned problems, we use two typical large-scale pretrained models to filter out passages: NEZHA (a masked language model) and NEZHA-Gen (a casual language model) (Wei et al., 2019).
- In order to take language features and difficulty level into account, we use new strategies and methods in passage collection, language model filtering and crowdsourced passage selection, which are different from LAMBADA.

We carry out an in-depth analysis on the built dataset, finding that the relations between target words and previous context ranges from lexical match, synonym, summary to commonsense reasoning. We conduct experiments on the built dataset to evaluate a range of state-of-the-art Chinese pretrained models, including the Chinese pretrained model PanGu-α with up to 200 billion parameters (Zeng et al., 2021), which achieves a top-1 accuracy of 12.1%, 45.2 points behind human

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3770–3778
November 7–11, 2021.

Dataset	Task	Data Collection	QA	Train	Development	Test	Language
CNN/Daily Mail	Entity Prediction	AC	✓	380,298/879,450	3,924/6,4835	3,198/53,182	EN
CBT	Entity Prediction	AC	✓	669,343	8,000	10,000	EN
LAMBADA	Word Prediction	RNNF+MC	✗	-	4,869	5,153	EN
WSC	Commonsense Reasoning	MC	✓	-	-	285	EN
WinoGrande	Commonsense Reasoning	MC	✓	-	-	44,000	EN
WPLC (ours)	Word Prediction	PLMF+MC	✗	-	4,827	4,474	ZH

Table 1: Comparison between our dataset and other datasets. AC: Automatically Chosen. RNNF: Filtering by RNN, MC: manual check. PLMF: Filtering by pretrained language models.

performance, indicating a large space for further research.

2 Related Work

CNN/Daily Mail (Hermann et al., 2015) uses an automatic method to create a large amount of instances of replacing entities with placeholders in news. Children's Book Test (CBT) (Felix et al., 2016) removes four types of words that are expected to be predicted by evaluated models and provides candidate choices for models. LAMBADA (Paperno et al., 2016) masks the last word in a target sentence and evaluates the ability of models in predicting the masked target words with broader context beyond target sentences in novels. Winograd Schema Challenge (WSC) (Levesque et al., 2012) and WinoGrande (Sakaguchi et al., 2020) defines a word selection task that focuses on solving commonsense problems in the form of coreference resolution. Details on the differences of Chinese WPLC from previous related datasets are shown in Table 1.

In Chinese, People Daily (PD) & Children's Fairy Tale (CFT) (Cui et al., 2016) corpus is the first cloze-style reading comprehension dataset in chinese. ChID (Zheng et al., 2019) offers an interesting task where words to be predicted are all idioms. CLUEWSC2020 (Xu et al., 2020), a Chinese version of WSC dataset, aims to test the ability of coreference resolution via word prediction. Significantly different from such Chinese datasets, our dataset is specifically developed for evaluating word prediction from long-range context.

3 Dataset Creation

3.1 Passage Collection

To diversify topics and domains, we collect raw data for the Chinese WPLC from 69,067 crawled novels with different topics (more details are shown in Table 2). The half of the crawled novels are used for training while the other half is used for extracting passages to build the development and test set. We automatically extract passages from raw data according to the following three rules:

Book Topics	Nums	Percentages (%)
Romance	22,292	32.3
Fantasy	12,190	17.6
Urban Supernatural	5,277	7.6
Comprehension	4,624	6.7
Rebirth	3,067	4.4
Science Fiction	3,023	4.4
Horror Suspense	2,162	3.1
Historical Military	1,868	2.7
Detective Mystery	1,379	2.0
Modern	1,252	1.8
Others	12,933	17.4
Total	69,067	100

Table 2: Topic distribution of crawled novels.

- As raw Chinese texts are not word-segmented, we use three different state-of-the-art Chinese word segmenters, PKUSEG (Luo et al., 2019), Jieba[1] and THULAC (Sun et al., 2016) to segment extracted passages. Only passages where the last word to be predicted can be consistently identified by the three segmenters are kept.
- If the last word is a stop word, the penultimate word will be considered as the target word as stop words are usually easily to be predicted. If the penultimate word is a stop word too, such passages will be discarded.
- We set the maximum length of a target word to 4, making the most difficult part of the task be to predict a Chinese idiom (four characters).
- The maximum length of passages is limited to 400 characters as long passages make word prediction more difficult even for humans.

3.2 Passage Filtering

Similar to LAMBADA (Paperno et al., 2016), we also use language models to filter out passages

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

where the target words (the last words) can be easily predicted by language models. But significantly different from LAMBADA, we use more powerful pretrained language models, instead of conventional or neural language models trained on relatively small data, to make our dataset challenging for state-of-the-art pretrained models.

We finetune NEZHA and NEZHA-Gen (Wei et al., 2019) on the training data which contain 8.7 billion words from 34,534 novels. We use two strategies to filter passages: (1) predicting the target word given a full passage (context + the target sentence that contains the target word) and (2) predicting the target word only given the target sentence. Such strategies are not only more rigorous than that used in LAMBADA but also consistent with the succeeding crowdsourcing step. Different combinations of the two pretrained models and strategies are used to filter passages.

In LAMBADA, a passage will be filtered out if the probability of the target word is greater than a preset threshold. Predefining an appropriate threshold is rather difficult, heavily depend on human experience. Thus, we use a different filtering method: any passages where the target word appears in the list of top-5 words predicted by either of the aforementioned two filtering strategies are discarded.

In addition to this, another difference is that we compute the ratio of the target word probabilities estimated given the full and target sentence by NEZHA-Gen as follows:

$$Ratio(w) = \frac{P(w|c, s_{\backslash w})}{P(w|s_{\backslash w})} \quad (1)$$

where $P(w|c, s_{\backslash w})$ is the probability of the target word w given the long-range context c plus the target sentence s excluding the target word w while $P(w|s_{\backslash w})$ is the probability of predicting w only given $s_{\backslash w}$. Higher ratios indicate that the target word can be more confidently predicted given the long-range context than the short-term context in the target sentence. Preference is given to passages with a ratio greater than the base e.

3.3 Crowdsourced Passage Selection

We hire over 100 crowdsourced workers to manually select passages from the left passages after the automatic passage collection and filtering procedure. For crowdsourced manual passage selection, we take 3 steps, similar to LAMBADA, where in the first two steps crowdsourced workers are asked to guess the missing target word given the entire passage excluding the target word.

In the third step, three different crowdsourced workers are asked to guess at most 3 target words per worker given the short-term context in the target sentence. If none of the manually predicted words are the target word, the passage is added to Chinese WPLC.

Particularly, in each step, workers are provided with the length of the target word to ease the guessing difficulty.

At last, we collect 9,301 passages, among which 4,827 passages from 17,266 novels are used as the development set while the remaining 4,474 passages from 17,267 novels are used as the test set. Table 3 provides the detailed statistics of the development and test set with respect to the target word length.

TWL	#Passages	#Avg tokens	#Avg sentences
1	408/354	117.7/119.7	3.7/4.3
2	3,904/3,670	130.7/136.1	3.6/4.3
3	260/236	130.3/137.1	3.7/4.4
4	255/214	128.2/127.2	3.8/4.0
total	4,827/4,474	129.5/134.5	3.6/4.3

Table 3: Statistics on the development/test set. TWL: the length of target words.

	Avg O	Avg DL	Avg DF
dev/test	1.3/1.3	72.8/74.1	81.5/83.8

Table 4: Statistics on lexical match on the dev/test set. Avg O: the average number of occurrences of target words in passage. Avg DF/DL: the average number of tokens between a target word and its first/last occurrence in context.

4 Dataset Analysis

4.1 Target Word Types

Figure 1 shows the distribution of the types of target words in Chinese WPLC. The majority of target words are common nouns (60.5%), followed by verbs (19.9%). Different from LAMBDA, Chinese WPLC contain 3.4% Chinese idioms (See the third example in Appendix Table 6). Chinese idioms increase the difficulty of word prediction for machine although they are widely used in human-written Chinese texts.

4.2 Linguistic Relations between Target Words and Long-Range Context

Inspired by Jing et al. (2019) and Paperno et al. (2016), we further analyze the linguistic relations

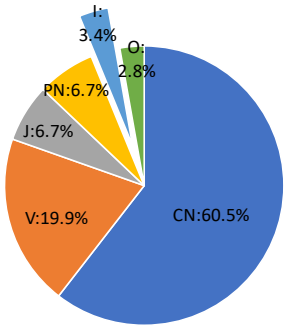

Figure 1: Target word type distribution. CN: common nouns. V: verbs. J: adjectives. PN: proper nouns. I: Chinese idioms. O: other.

between target words and long-term context in passages. We sample 100 examples from the development set and find four linguistic relations: lexical match, synonym, summary, reasoning as shown in Appendix Table 6. Lexical match, indicating that the target word has also occurred in context, accounts for 64%. However, lexical match does not mean that the target word can be easily predicted as further statistics in Table 4 disclose that the distance between the target word and its first/last apperance in context is very long, ranging from over 70 to 80 tokens. Synonym, suggesting that a word or phrase with similar meaning to the target word occurrs in context, accounts for 15%. A more difficult phenomenon is to summarize the given passage to predict the target word, which accounts for 8% of the sampled data. The left samples need to conduct reasoning over context while the target word has not been explicitly mentioned in context at all.

5 Experiments

We carried out experiments with a range of state-of-the-art pretrained language models on Chinese WPLC. As BERT-large and the last layer of RoBERTa-large are currently not available for Chinese, results of these two models are not provided. Top-1 and Top-3 accuracy are reported.

5.1 Baseline Models

In addition to BERT (Devlin et al., 2019), we also evaluated the following pre-trained language models on the dataset.

- **ALBERT:** ALBERT (Lan et al., 2020) is a lite BERT with fewer parameters but more powerful performance.
- **RoBERTa:** RoBERTa (Liu et al., 2019) is a stronger BERT without the next sentence prediction loss.
- **MacBERT:** MacBERT (Cui et al., 2020) is a Chinese BERT that uses similar words for the masking purpose.
- **CPM:** CPM(Zhang et al., 2020) is a Chinese GPT-2 (Radford et al., 2019) with 2.6 billion parameters.
- **PanGu-α:** PanGu-α (Zeng et al., 2021) is a Chinese pre-trained casual language model with up to 200 billion parameters. The version that we used in experiments has 13 billion parameters.

5.2 Experimental Setup

All baselines were tested using their default hyper-parameters, including BERT[2], ALBERT[3], RoBERTa[2], MacBERT[4], CPM[5] and PanGu-α[6]. For causal language models, beam-search was used to generate top-3 words and the number of generation steps was the length of the target word. For masked language models, we downloaded a whole word mask version and selected top-3 words in the masked positions as predicted target words.

5.3 Human Evaluation

In order to assess human performance on Chinese WPLC, we hired another 4 crowdsourced workers to perform word guessing on 1000 samples randomly chosen from the development and test set (500 each). Each worker is asked to guess 3 words and the first word is considered as the most probable word guessed by worker.

5.4 Results

Table 5 presents the results of the models on the development and the test data. Note that the scores of NEZHA and NEZHA-Gen are 0 since they are used to filter passages in Section 3.2.

Pretrained Models vs. Human: All state-of-the-art pretrained models perform much worse than human on this task. PanGu-α achieves a top-1 accuracy of 12.1%, the highest prediction accuracy among all pretrained models, which, however, is

[2]https://github.com/ymcui/Chinese-BERT-wwm
[3]https://github.com/google-research/ALBERT
[4]https://github.com/ymcui/MacBERT
[5]https://huggingface.co/mymusise
[6]https://git.openi.org.cn/PCL-Platform.Intelligence/PanGu-Alpha

	Top-1	Top-3
Nezha-Gen	0/0	0/0
Nezha	0/0	0/0
Human	57.3	66.4
Casual Language Models		
CPM	0.6/0.5	1.5/1.5
CPM-kd	1.2/0.9	2.9/2.4
PanGu-α	12.7/12.1	-/-
Masked Language Models		
BERT-base	7.3/6.3	10.1/8.9
RoBERTa-base	6.5/5.7	9.8/8.9
MacBERT-large	6.8/7.5	10.6/10.5
ALBERT-xxlarge	4.5/3.8	6.5/5.4

Table 5: Top-1 and Top-3 accuracy (%) results of models and human on the development/test of Chinese WPLC. CPM-kd: knowledge distillated (Geoffrey et al., 2015) CPM.

45.2 points behind human performance (57.3%). We find that knowledge distillation helps in CPM-large achieve a gain of 0.4 to 1.4 percentage points.

Masked Language Models (MLMs) vs. Casual Language Models (CLMs): MLMs (BERT-like) are slightly better than CLMs (next token prediction) in Table 5. The reasons may be two-fold. First, since MLMs are bidirectional, they can use extra information after target words, such as stop words and punctuations, to predict target words. Second, we used stronger NEZHA-Gen to filter out passages in dataset creation, which may make the remaining passages difficult for other CLMs.

5.5 Analysis on PanGu-α and Human Prediction

We analyzed 100 randomly sampled passages from the development set to compare PanGu-α with crowdsourced workers. One difference between human and models on word prediction on Chinese WPLC is that human workers can use the length of a target word as auxiliary information to predict target word while current models cannot use such information. We find that 14% of predicted words by PanGu-α are completely correct and 22% are almost correct (See the first and second example in Appendix Table 7). There are also 11% of examples where target words predicted by PanGu-α are similar to the ground-truth target words (See the third example in Appendix Table 7).

We also analyzed 100 sampled passages with correct word predictions by human workers and PanGu-α. We find that 75% of these human predictions are lexical match and 7% are synonym. The type of summary accounts for only 4% of passages while the left 14% are reasoning. For PanGu-α, 71% of predictions are lexical match followed by reasoning which accounts for 23%. There are also 4% of synonym, followed by summary, which accounts for 2%. Lexical match is the easiest type for both human and models. Even the target words of reasoning-type word prediction have not been explicitly mentioned in context at all, we find that both human and models can do better than they do in the other two types (i.e., synonym and summary).

6 Conclusions

In this paper, we have presented the Chinese WPLC, a Chinese word prediction dataset created from over 69K novels to examine the ability of pretrained language models on long-term context modeling. We employ both automatic and manual selection strategies to keep passages where target words can be only predicted from long-term context beyond target sentences and it is difficult for pretrained language model to predict target words. Experiments with a range of state-of-the-art pretrained language models and in-depth analyse demonstrate that the created dataset is a very challenging testbed even for the very large Chinese pretrained PanGu-α, covering a variety of linguistic phenomena (e.g., lexical match, synonym, summary and reasoning).

Acknowledgements

The present research was supported by Huawei. We would like to thank the anonymous reviewers for their insightful comments. The corresponding author is Deyi Xiong (dyxiong@tju.edu.cn).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computa-*

tional Linguistics: EMNLP 2020, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. Dataset for the first evaluation on Chinese machine reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for Chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hill Felix, Bordes Antoine, Chopra Sumit, and Weston Jason. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Hinton Geoffrey, Vinyals Oriol, and Dean Jeffrey. 2015. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1693–1701. MIT Press.

Yimin Jing, Deyi Xiong, and Zhen Yan. 2019. BiPaR: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2452–2462, Hong Kong, China. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. 2019. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. *CoRR*, abs/1906.11455.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Austin, Texas. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WINOGRANDE: an adversarial winograd schema challenge at scale. *In AAAI*.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. THULAC: An efficient lexical analyzer for chinese.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. NEZHA: Neural Contextualized Representation for Chinese Language Understanding. *arXiv preprint arXiv:1909.00204*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and

Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation. *CoRR*, abs/2104.12369.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. 2020. CPM: A Large-scale Generative Chinese Pre-trained Language Model.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.

A Appendix

Relations	Example	%
Lexical match	**Passage:** 在一小时的时间里他一直在睡觉。科伦巴的小机场非常潮湿，那儿聚集着一群等候去圣克鲁斯的玻利维亚人。他们个个带着大包小包的圣诞礼物。他叫的那位出租车司机不懂一句英语，但这没关系。内特指给他看旅游手册上的"皇宫饭店"几个字，他坐上这辆又旧又脏的出租车离开了<mask><mask>。 He had been sleeping during this hour. The small airport in Corumba was very humid, and there was a crowd of Bolivians waiting to head for Santa Cruz.They all carry bags of Christmas presents . The taxi driver he called didn't understand a word of English, but it didn't matter. Nate pointed out the words "Palace Hotel" in the travel brochure. He got in this old and dirty taxi and left the <mask>. **Target word:** 机场 / airport	64
Synonym	**Passage:** 守殿的大太监名叫过业大，人称大公公。国藩与大公公打声招呼后，便端坐在养性殿候驾。一坐整整两个时辰，时至正午，尚不见召，国藩心中犯疑，请大公公打听。一会儿，大公公告诉他："皇上今天不来了，明天在养心殿<mask><mask>。" The eunuch who guarded the temple was called Guoyeda, commonly known as "the Grand Eunuch". Having greeted the Grand Eunuch, Guofan sat in the Hall of Mental Cultivation waiting for the emperor's coming. Guofan sat for two hours until noon, but still didn't get called. He got bewildered, so asked the Grand Eunuch to inquire about this. After a while, the Grand Eunuch told him: "The emperor is not coming today, but will <mask> you at the Hall of Mental Cultivation tomorrow." **Target word:** 召见/ summon	15
Summary	**Passage:** 健康的红色会让他们的无限遐想通过努力逐渐转变成为现实，而遗憾的是那些没有自制力的红色却疏于行动，很多梦想最终堕落为空想。因此，与其说堂吉珂德是西班牙的最后一位骑士，莫如说他是超级富于幻想的红色代表人物。当然，如果红色不停地空想，再加上夸夸其谈，一不小心，变成"<mask><mask><mask><mask>"。 The healthy red can make their infinite daydream changing gradually to a reality through efforts. However, it is a pity that those red who have no self-control failed to take actions, and many dreams eventually degenerates into fantasies. Thus, Don Quixote is not so much the last knight of Spain as a super fanciful representative of the red. There is no doubt that if the red cannot stop indulging in fantasy, even in some magniloquence, it will turn into <mask><mask><mask><mask>"easily. **Target word:** 纸上谈兵 / an idea on paper	8
Reasoning	**Passage:** 孟飞酝酿了半天硬是没叫出爸和妈，苏蓝为孟飞解围说："他第一次见你们，一时半会还不习惯。"她妈妈非常宽容地说："小伙子第一次总是很难说出口的，结了婚就慢慢习惯了。"孟飞一听窃喜，这话表示她妈妈已经默许了他这位<mask><mask>。 Meng Fei had been brewing for a long time but did not call out father and mother in the end. Su Lan helped him out and said, "It's the first time he has met you, so he doesn't get quite used to it in such a short time." Her mother said very tolerantly: "It has always been hard for a young man to say this for the first time, but you'll get used to it after you got married." Meng Fei was secretly pleased on hearing that, which indicated that her mother had acquiesced in him as a <mask><mask><mask>. **Target word:** 女婿 / son-in-law	13

Table 6: Linguistic relations between target words and long-term context. Each "<mask>" represents a single Chinese character.

Example
Passage: 刚才你也都看到了，在发病时候的她，完全就把我这个哥哥当成毒蛇猛兽一般，她非常的排斥我，不愿意见我，所以，我爸妈就借着公司事务，在那段时间把我调离国外去处理事情，等我回来，她已经被送到了……那个医院，我去看她，她也从来都是<mask><mask><mask><mask>…… As you have seen just now, during the attack, she completely regarded me, his brother, as a venomous serpent and wild beast. She ostracized me very much and was reluctant to see me. Therefore, under the guise of company affairs, my parents sent me abroad to deal with the business. When I came back, she had been sent to...that hospital. I went to see her, but she had always been <mask><mask>..... **PanGu-α:** 不愿意见我 / reluctant to see me **Target word / Human:** 避而不见 / evading me
Passage: 老爷子年迈了,身体也没以前硬朗,可还是不服输的性子,不过,我却越来越察觉,他对于财富,已没有当年那般热衷,陆氏旗下有多少企业,有多少资产,于他,也只是一纸符号,人老了,最盼望的还是一家团聚!有时间,你多给家桓旁敲侧击下,让他早点回来,不仅是陆氏等他,还有<mask><mask><mask>。 The old man is getting older and his body is also not as strong as before. But he still has an unyielding personality. However, I have become more and more aware that he is no longer as enthusiastic about wealth as he used to be in the past. No matter how many enterprises and assets are owned by the Lu's group, it's just a paper of symbols for him. When people are old, what they most look forward to is family reunion! When you have time, you could insinuate Jiahuan that he should come back early. Not only is Lu waiting for him, but also <mask><mask>. **PanGu-α:** 你老爷子 / your old man **Target word / Human:** 老爷子 / old man
Passage: 有时对方正急需,又不肯对你明言,或故意表示无此急需,你如得知情形,更应尽力帮忙,并且不能有丝毫得意的样子,一面使他感觉受之有愧,一面又使他有知己之感。寸金之遇,一饭之恩,可以使他终生铭记。日后如有所需,他必奋身图报。即使你无所需,他一朝否极泰来,也绝不会忘了你这个<mask><mask>! Sometimes one is in desperate need of you, but would not tell you clearly, or deliberately indicate that there is no urgent need. If you know this situation, you should try your best to help, and cannot show any complacency. On the one hand, it would make him shameful for receiving it and give him the feeling of having a new confidant on the other hand. The encounter of an inch of gold and the grace of a meal can make him remember for life. And if you need help later, he will go out of his way to help you. Even if you don't need it, after a storm comes a calm, he will not forget you who is his <mask>! **PanGu-α:** 朋友 / friend **Target word / Human:** 知己 / confidant

Table 7: Examples with predicted target words from PanGu-α and humans.