BBT-Fin: Comprehensive Construction of Chinese Financial Domain Pre-trained Language Model, Corpus and Benchmark

Dakuan Lu[1], **Jiaqing Liang**[2], **Yipei Xu**[1], **Qianyu He**[1],
Yipeng Geng[3], **Mengkun Han**[3], **Yingsi Xin**[3], **Hengkui Wu**[3]*, **Yanghua Xiao**[1]*
[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2]School of Data Science, Fudan University
[3]SuperSymmetry Technologies
{ludakuan1234, l.j.q.light, xuyipei000, abbey4799}@gmail.com,
{ypgeng, mkhan, ysxin, hkwu}@ssymmetry.com, shawyh@fudan.edu.cn

Abstract

To advance Chinese financial natural language processing (NLP), we introduce BBT-FinT5, a new Chinese financial pre-training language model based on the T5 model. To support this effort, we have built BBT-FinCorpus, a large-scale financial corpus with approximately 300GB of raw text from four different sources. In general domain NLP, comprehensive benchmarks like GLUE and SuperGLUE have driven significant advancements in language model pre-training by enabling head-to-head comparisons among models. Drawing inspiration from these benchmarks, we propose BBT-CFLEB, a Chinese Financial Language understanding and generation Evaluation Benchmark, which includes six datasets covering both understanding and generation tasks. Our aim is to facilitate research in the development of NLP within the Chinese financial domain. Our model, corpus and benchmark are released at https://github.com/ssymmetry/BBT-FinCUGE-Applications. Our work belongs to the Big Bang Transformer (BBT), a large-scale pre-trained language model project.

1 Introduction

Pre-trained language models(PLMs), such as BERT (Devlin et al., 2018) and T5 (Raffel et al., 2019), have led to great performance boosts across many NLP tasks. Despite the excellent performance of pre-trained language models (PLMs) on a large number of NLP tasks, their performance is often affected when applied to domain-specific texts that exhibit significant differences from general text in terms of word usage, syntax, and writing style (Gururangan et al., 2020; Gu et al., 2021). To address this issue, Gururangan et al. (2020) proposed that continuing to pre-train a general PLM on target domain corpora and task-relevant texts can effectively improve its performance on domain-specific tasks, while Gu et al. (2021) further suggested that pre-training domain-specific PLMs from scratch with a sufficiently large corpus can achieve even better domain-specific performance. Inspired by these studies, domain-specific pre-trained language models have emerged in some domains, such as BioBERT(Peng et al., 2019a) and PubMedBERT (Gu et al., 2021) in the biomedicine field, which have been utilized for practical tasks like entity and relation extraction.

We collect all existing NLP competition tasks and academic datasets related to finance on the Chinese internet and summarized them in Table 2, revealing a growing demand for NLP capabilities in finance, particularly in information extraction and sentiment analysis. To meet these demands and improve the overall level of Chinese financial NLP, several companies have already developed and released Chinese financial pre-trained language models, such as FinBERT (Hou et al., 2020) and Mengzi-BERT-base-fin (Zhang et al., 2021). However, these models are based on the BERT-base model, have a single architecture type, and a parameter count (around 110 million) that is outdated and unable to meet the increasing demand for NLP capabilities in this field. Therefore, we propose FinT5, the largest Chinese financial pre-trained language model to date, based on the advanced T5 architecture, with 220 million parameters for the base version and 1 billion for the large version.

Furthermore, NLP tasks in the financial industry focus primarily on information extraction, requiring models with high entity knowledge understanding and memorization capabilities. Although studies have shown that pre-trained PLMs on large-scale corpora already have some entity knowledge understanding and memorization capabilities, there are still some shortcomings. To address this issue, many studies have used knowledge-enhanced pre-training methods to improve PLMs' understanding and memorization of entity knowledge. However,

*Corresponding author.

these methods mostly target BERT-like models and lack strategies designed for T5 models. To improve T5's performance on financial NLP tasks, we propose a concise knowledge-enhanced pre-training method based on the T5 model's text-to-text paradigm.

In addition, another challenge faced by Chinese financial NLP is the lack of corpus. The scale and diversity of corpora play an essential role in language model pre-training (Xu et al., 2020; Raffel et al., 2019; Gao et al., 2020). However, existing Chinese financial corpora are small in scale, poor in diversity and not open, as can be shown in Table 1. To solve this problem, we first need to determine the text types that a qualified Chinese financial corpus needs to cover. To this end, we first collected almost all existing Chinese financial NLP tasks and summarized their text sources, as shown in the Table 2. According to the source distribution of these tasks, we have determined the range of text types we need to collect. As a result, we collect and release a large-scale Chinese financial corpus named BBT-FinCorpus with about 300 GB raw text, which consists of five different sources to enhance its diversity covering most text sources of Chinese financial NLP tasks.

The widespread use of benchmark evaluations is a key driving force that has greatly improved and rapidly iterated PLMs. These evaluations use a single score to assess model performance across multiple tasks, enabling direct and comprehensive comparisons between pre-trained language models. Existing English PLMs use the general benchmark evaluations GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019), while the general benchmark evaluation for Chinese PLMs is CLUE (Xu et al., 2020). Almost all PLMs participate in these evaluations to compare their performance with other models. However, there is no publicly available benchmark for Chinese financial NLP, which makes it difficult to compare existing pre-trained language models on different task sets and hinders the rapid improvement of PLM performance in the Chinese financial domain.

To address this issue and promote research in the financial domain, we propose CFLEB, the **C**hinese **F**inancial **L**anguage Understanding and Generation **E**valuation **B**enchmark, consisting of six datasets covering language understanding and generation tasks. These datasets encompass a diverse range of text genres, dataset sizes, and levels of difficulty, and more importantly, emphasize challenges that arise in real-world scenarios.

Our contributions are summarized as follows:

- We introduce BBT-FinT5, a state-of-the-art financial Chinese PLM with large-scale parameters and knowledge-enhanced pre-training.
- We provide BBT-FinCorpus, a comprehensive and diverse financial Chinese corpus.
- We propose BBT-CFLEB, a benchmark for evaluating Chinese language understanding and generation in the financial domain.

2 Related Work

2.1 Domain-specific PLMs and Corpora

PLMs have achieved state-of-the-art performance in many NLP tasks (Devlin et al., 2018; Raffel et al., 2019; Liu et al., 2019). However, when applied to domain-specific tasks, models pre-trained on general corpora often produce unsatisfactory results due to the difference in word distribution from general to specific domains (Gururangan et al., 2020; Gu et al., 2021). To better adapt a language model to a target domain, pre-training on the corpus of the target domain is proposed (Gururangan et al., 2020). For domains with abundant unlabeled text, such as biomedicine, pre-training from scratch results in substantial gains over continual pre-training of general-domain language models (Gu et al., 2021). Consequently, many domain-specific PLMs have been proposed and pre-trained on their respective corpora.

In the field of financial NLP, domain-specific pre-trained language models (PLMs) have demonstrated their superiority over general-domain PLMs. For instance, Araci (2019) and Yang et al. (2020) pre-trained BERT on English finance news and communications, respectively, and outperformed competitive baselines on financial sentiment analysis tasks. In the context of Chinese financial NLP, Hou et al. (2020) pre-trained BERT on Chinese financial news, analysis reports, company announcements, and encyclopedias, and evaluated it on news classification, sentiment analysis, and named entity recognition tasks. Furthermore, Zhang et al. (2021) pre-trained the Chinese PLM Mengzi on a 20GB financial corpus and demonstrated its effectiveness on multiple downstream tasks.

Table 1 summarizes the characteristics of typical PLMs and their corpora in the financial domain. It

can be observed that both the scale of our model and corpus exceed existing works.

2.2 Knowledge Enhanced Pre-training

Although PLMs can acquire rich linguistic knowledge from pretraining on large-scale corpora, many studies have shown that PLMs still have shortcomings in entity knowledge understanding and memory, as the distribution of entity knowledge in unfiltered corpora is sparse and long-tailed (Yang et al., 2021). Therefore, PLMs can benefit from knowledge-enhanced pretraining methods that strengthen entity knowledge understanding and memory.

For example, Ernie (Sun et al., 2019) is designed to learn language representation enhanced by knowledge masking strategies, which includes entity-level masking and phrase-level masking. The disadvantage of this approach is that it can only help the model better learn existing entity knowledge from the corpus, without addressing the issues of sparse and long-tailed distribution of entity knowledge in the corpus.

Ernie 3.0, introduced by Sun et al. (2021), incorporates the universal knowledge-text prediction (UKTP) task. This task involves a pair of triples from a knowledge graph and their corresponding sentences from an encyclopedia, where either the relation in the triple or the words in the sentence are randomly masked. In order to predict the relation in the triple, the model must identify the head and tail entities mentioned in the sentence, and determine the semantic relationship between them.

The limitation of this approach is that it only masks the relation in the triple and not the entities, which can hinder the learning of entity representations. Moreover, distant supervision has a certain amount of noise, which means that the relation in the triple may not necessarily appear in the sentence (Smirnova and Cudré-Mauroux, 2018). Therefore, only masking the relation and predicting it can have a strong negative impact on the model. Although the above methods have made some progress, they are all designed for the BERT-like model.

To our knowledge, there is currently a gap in knowledge enhancement pre-training methods available for T5-like models.

2.3 Domain-specific NLP Benchmarks

Various domain-specific NLP benchmarks have been proposed to compare the ability of different methods in modeling text from specific domains in a fair manner. The BLUE benchmark (Peng et al., 2019b) evaluates the ability of models in biomedical text mining through five tasks. The BLURB benchmark (Gu et al., 2021) further focuses on clinical domains by removing two unrelated tasks and includes a wider range of biomedical applications. Despite these efforts, a comprehensive set of benchmark tasks for training, evaluating, and analyzing financial PLMs is still largely unexplored. Currently, the FLUE (Shah et al., 2022) is the only benchmark for the financial domain, consisting of five tasks specifically designed for English financial text. However, we are the first to construct a comprehensive set of benchmarks for Chinese financial text, covering a range of language understanding and generation tasks that differ from previous works.

3 The Corpus: BBT-FinCorpus

We build FinCorpus, the biggest corpus of Chinese financial domain to get a superior pre-trained language model. Section 3.1 covers how we decided on the corpus contents. We collected, refined and sorted the corpus to finally obtain the FinCorpus, as elaborated in Section 3.3.

3.1 Coverage Confirmation of the Corpus

We believe that, since the purpose of domain pretraining is to help models better understand domain texts and perform domain tasks more effectively, it is essential to observe the text distribution of domain tasks to determine the coverage of the corpus. The domain corpus should cover the text sources of domain tasks as much as possible to enhance the model's understanding of the tasks. To this end, we first collected almost all Chinese financial NLP task datasets available on the Chinese internet in recent years, including several datasets used in this study, and their text sources, as shown in Table 2.

It can be seen that the text sources of these financial NLP datasets are mainly concentrated in financial news, company announcements, research reports, and social media. For financial news, we chose the largest financial news websites on the Chinese Internet for crawling, namely Sina Finance [1], Tencent Finance [2], Phoenix Finance [3],

[1] https://finance.sina.com.cn/
[2] https://new.qq.com/ch/finance/
[3] https://finance.ifeng.com/

PLM	Size	Corpus Size	Corpus Sources
FinBERT(Araci, 2019)	110M	29M words	News filtered by financial keywords
FinBERT(Yang et al., 2020)	110M	4.9B tokens	Corporate Reports, Earnings Call Transcripts, Analyst Reports
FinBERT (Hou et al., 2020)	110M	3B tokens	News, Analyse reports, Company announcements and Encyclopedias
Mengzi-BERT-base-fin (Zhang et al., 2021)	110M	20GB file	News, Analyse reports, Company announcements
BBT-FinT5 (ours)	220M, 1B	80B tokens	Corporate Reports, Analyst Reports, Social media and Financial News

Table 1: Typical financial PLMs and their corpora.

Dataset	Text Source	Open State	Practicality
DuEE-fin (Han et al., 2022)	Financial news, Company announcement	Yes	High
FinRE (Li et al., 2019)	Financial news	Yes	High
Announcement information extraction (Tianchi, 2018)	Company announcement	Yes	High
Discovery of new entities in Internet finance (Datafountain, 2019)	Social media	Unspecified	Low
Announcement information extraction (Biendata, 2019)	Company announcement	Unspecified	High
Construction of financial knowledge graph (Biendata, 2020b)	Analyse report	Unspecified	Medium
Event causality extraction (Biendata, 2021)	Financial news	Unspecified	Low
Financial NL2SQL (Biendata, 2022a)	Data query sentence	Unspecified	Medium
Few-shot event extraction (Biendata, 2022b)	Financial news	Unspecified	Medium
Few-shot event extraction (Biendata, 2020a)	Financial news	Unspecified	Medium
FinNL (ours)	Financial news	Yes	High
FinNA (ours)	Financial news	Yes	High
FinFE (ours)	Social media	Yes	High
FinNSP (ours)	Social media	Yes	High

Table 2: Chinese financial datasets we collected, with their open source status and practicality scores

36Kr [4]and Huxiu [5]. For company announcements and research reports, we chose Eastmoney [6] for crawling. For social media, we chose the two largest financial social media platforms on the Chinese Internet, Guba [7] and Xueqiu [8], for crawling.

3.2 Crawling and Filtering of the Corpus

We used a proxy-based distributed crawler to crawl public web pages. We filtered the web pages using a series of rules (Raffel et al., 2019; Yuan et al., 2021).

[4]https://36kr.com/
[5]https://www.huxiu.com/
[6]https://www.eastmoney.com/
[7]https://guba.eastmoney.com/
[8]https://xueqiu.com/

3.3 Description of the Corpus

After crawling, cleaning, and processing, we obtained the FinCorpus, a large-scale Chinese financial domain corpus that contains four types of language materials:

- **Corporate announcements.** These are the announcements released by all listed companies in China over the past twenty years. The original data is in PDF format, with a total size of about 2TB. Using a PDF parser, we converted the PDF files into text files, resulting in a total size of 105GB.

- **Research reports.** These are research reports issued by investment institutions such as securities firms and investment banks on macroeconomic issues, sectors, industries,

and individual stocks, analyzing the current status and future development trends of the research object. The original data is in PDF format, with a total size of about 1TB. After conversion, the total size of the resulting text files is about 11GB.

- **Financial news.** These are the financial news articles from the past five years crawled from websites including Sina Finance, Tencent Finance, Phoenix Finance, 36Kr, and Huxiu. After cleaning, the total size of the resulting text files is about 20GB.

- **Social media.** These are the posts from all stockholders and bloggers published on stock bar and Xueqiu website over the past twenty years. After cleaning, the total size of the resulting text is about 120GB.

The corpus from the above five sources basically covers all types of texts in the common Chinese financial NLP.

4 The Large PLM: BBT-FinT5

To enhance the performance of the Chinese financial NLP baseline and foster the growth of the open-source community in this domain, we introduce the FinT5 model. This model's architecture and pre-training tasks are consistent with the T5 (Raffel et al., 2019) model and are pre-trained on BBT-FinCorpus (refer to Section 3). We chose this model for its robust performance on many general benchmarks and compatibility with understanding and generating tasks based on the text-to-text paradigm, which facilitates transfer learning. Our experiments demonstrate that the FinT5 model significantly outperforms T5 trained on the general corpus.

In this section, we first describe the architecture and pre-training task of the T5 model. Then we outline the pre-training acceleration method based on DeepSpeed, and finally introduce the knowledge enhancement pre-training method that we propose for the T5 model, which is based on triple masking.

4.1 Pre-training Model Architecture and Task

Raffel et al. (2019) model all NLP tasks in a text-to-text format which enable the use of a unified network architecture, training approach, and loss function to handle all NLP tasks, promoting transfer learning in the NLP field. Building upon this, they conducted a series of comparative experiments and chose to develop a large-scale PLM, T5, based on an encoder-decoder architecture and pre-trained using MLM. Specifically, T5 utilizes the span mask method proposed by SpanBERT (Joshi et al., 2020), randomly masking 15% contiguous spans within a sentence rather than independent tokens.

4.2 Pre-training Acceleration

We use the optimizer state parallelism and gradient parallelism implemented by DeepSpeed (Rasley et al., 2020) to accelerate the pre-training process. In particular, we found that using the BFLOAT16 (Kalamkar et al., 2019) half-precision floating-point format for optimization can effectively solve the problem of gradient overflow that occurs in the training process with FP16 half-precision floating-point format, without the need to repeatedly adjust gradient scaling coefficients and other hyperparameters. Kalamkar et al. (2019) pointed out that in the training of deep neural networks, the value range (i.e., exponent range) of the floating-point numbers used to represent each parameter in the network is more important for training stability and performance than their mantissa precision. Therefore, the BFLOAT16 format uses the same eight-bit exponent as the FP32 format to represent the same exponent range as the FP32 format, at the cost of having three fewer mantissa bits than the FP16 format. Extensive experiments have shown that this trade-off makes the BFLOAT16 format as fast and memory-efficient as the FP16 format while having training stability and performance close to that of the FP32 format.

4.3 Knowledge Enhancement Pre-training Method Based on Triple Masking

We propose a knowledge enhancement pre-training method based on triple masking (KETM).

First, for each triple in the knowledge graph, we use the distant supervision algorithm to obtain sentences corresponding to it. Specifically, for a knowledge triple (head entity, relation, tail entity), if there is a sentence in the encyclopedia that contains both the head and tail entities, we consider this sentence to contain the knowledge described by this triple.

Next, for a sentence and its contained triple, we concatenate the triple at the beginning of the sentence. For the triple part, we randomly mask one element, and for the sentence part, we randomly mask 15% of a random-length span. Finally,

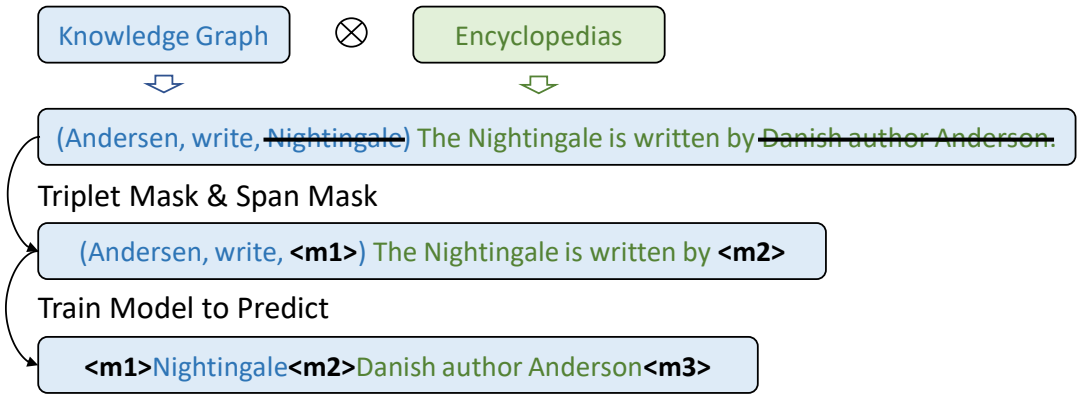

Figure 1: Knowledge enhancement pre-training method based on triple masking (KETM)

we input the masked triple and sentence into the model and require the model to predict the masked element, as shown in the Figure 1. The model is trained to fill the masked element in the triple based on the two unmasked elements in the triple and the partially masked sentence, which helps the model better understand and memorize entity-related knowledge.

5 The Benchmark: BBT-CFLEB

In this section, we first describe the method used for selecting tasks for the benchmark. We then introduce the selected tasks and the three leaderboards, each of which is composed of different tasks.

5.1 Task Selection

We propose that for domain-specific NLP evaluation benchmarks, special attention should be paid to their practicality, especially for the financially valuable field, to better reflect the model's ability in practice. Therefore, we use a practicality score to measure the practicality of the tasks we collect. Specifically, we invited financial experts to evaluate the practicality of each task and gave a low, medium, or high practicality rating, only selecting tasks with a high practicality rating as candidate tasks. In addition, we only kept tasks with a clear open-source statement as candidate tasks. Finally, we selected six tasks for BBT-CFLEB in Table 2.

5.2 Task Introduction

CFLEB includes six tasks in total, consisting of two language generation tasks and four language understanding tasks. These tasks are as follows:

- FinNL, a financial news classification dataset. Given financial news articles, the model needs to classify them into up to 15 possible categories, with evaluation measured by F1-Score. The training set contains 8,000 articles, the validation set contains 1,000 articles, and the test set contains 1,000 articles.

- FinNA, a financial news summarization dataset. Given financial news articles, the model needs to generate a summary, with evaluation measured by Rouge (Lin, 2004). The training set contains 24,000 articles, the validation set contains 3,000 articles, and the test set contains 3,000 articles.

- FinRE, a financial news relation extraction dataset. Given financial news articles and head-tail entity pairs, the model needs to classify the relation between entity pairs into up to 44 categories, including the null relation, with evaluation measured by F1-Score. The training set contains 7,454 articles, the validation set contains 1,489 articles, and the test set contains 3,727 articles.

- FinFE, a financial social media text sentiment classification dataset. Given financial social media text, the model needs to classify the sentiment of the text into negative-neutral-positive categories, with evaluation measured by accuracy. The training set contains 8,000 articles, the validation set contains 1,000 articles, and the test set contains 1,000 articles.

- FinQA, a financial news announcement event question-answering dataset, derived from the DuEE-fin (Han et al., 2022) dataset. Given

Task Name	Introduction	Data	Evaluation
FinNL	Multi-label classification of financial news	8000/1000/1000	F1-score
FinNA	Generation of summaries for financial news	24000/3000/3000	Rouge
FinRE	Entity relation classification for financial news	7454/1489/3727	F1-score
FinFE	Sentiment classification of financial social media text	8000/1000/1000	Accuracy
FinQA	Question-answering for financial news/events	16000/2000/2000	F1-score
FinNSP	Detection of negative messages and entities in financial news	4800/600/600	F1-score

Table 3: Summary of CFLEB tasks.

financial news or announcement text and a question related to an event mentioned in the text, the model needs to generate an answer to the question based on the text, with evaluation measured by F1-Score. The training set contains 16,000 articles, the validation set contains 2,000 articles, and the test set contains 2,000 articles.

- FinNSP, a financial negative news and its subject determination dataset. Given financial news or social media text and entities mentioned in the text, the model needs to determine if the text contains negative news related to any entity and identify which entity is the subject of the negative news, with evaluation measured by F1-Score. The training set contains 4,800 articles, the validation set contains 600 articles, and the test set contains 600 articles.

5.3 Leaderboard Introduction

We have organized the tasks into multiple leaderboards according to different ability requirements (Xu et al., 2020), so that researchers can observe the model's ability rankings from different perspectives. The leaderboards of FinCUGE are as follows:

- Overall leaderboard: includes all six tasks.
- Understanding ability leaderboard: includes four language comprehension tasks, FinNL, FinRE, FinFE, and FinNSP.
- Generation ability leaderboard: includes two language generation tasks, FinNA and FinQA.

6 Experiments

In this section, we first introduces the basic settings of the experiment, including the basic information of the PLMs involved in the comparison and the processing format of the tasks in the evaluation benchmark. Then we conduct sufficient experimental and comparative analysis to validate the effectiveness of the proposed model and method.

6.1 Experiments Setup

6.1.1 Pre-trained Language Models

The models participating in the comparative experiment of this section include:

- **GPT2-base** (Zhao et al., 2019). A Chinese GPT2 released by Zhao et al. (2019). Pre-trained using the general corpus CLUECorpusSmall (Xu et al., 2020).
- **T5-base** (Zhao et al., 2019). A Chinese T5 released by Zhao et al. (2019). Pre-trained using the general corpus CLUECorpusSmall (Xu et al., 2020).
- **FinBERT** (Hou et al., 2020). A Chinese BERT for the financial domain released by Hou et al. (2020).
- **Mengzi-BERT-base-fin** (Zhang et al., 2021). A Chinese BERT for the financial domain released by Zhang et al. (2021).
- **FinT5-base**. Our Chinese pre-trained language model for the financial domain, pre-trained on our financial corpus, FinCorpus. Its model architecture, parameter size, and

PLMs	FinFE	FinNL	FinNSP	FinRE	Un.Avg.	FinNA	FinQA	Ge.Avg.	Avg.
GPT2-base	79.05	84.09	91.30	36.37	72.70	44.19	75.22	59.71	68.37
T5-base	79.40	87.48	**95.43**	54.93	79.56	48.54	83.58	66.06	74.89
FinBERT-base	79.45	84.69	69.01	55.33	72.37	-	-	-	-
Mengzi-BERT-base-fin	79.50	85.88	71.72	58.25	73.59	-	-	-	-
BBT-FinT5-base	80.19	87.55	94.50	60.62	80.21	50.06	84.82	67.44	76.29
BBT-FinT5-base-KE	79.43	87.77	95.05	61.79	80.26	51.36	85.66	68.51	76.84
BBT-FinT5-large	**80.24**	**88.44**	94.54	**61.88**	**81.78**	**51.42**	**85.95**	**68.69**	**77.07**

Table 4: Results of BBT-CFLEB from different PLMs.

pre-training hyperparameters are the same as T5-v1.1-base.

- **FinT5-base-KE**. Knowledge-enhanced version of FinT5-base, enhanced by KETM method using CN-DBPedia (Xu et al., 2017) knowledge graph.

- **FinT5-large**. Our proposed Chinese pre-trained language model for the financial domain, with a total of about 1 billion model parameters, and the pre-training hyperparameters are the same as T5-base.

6.1.2 Fine-tuning

For generative models (GPT, T5), we evaluated all six datasets by modeling all tasks as text-to-text. For BERT-based models, we evaluated them on four language understanding tasks: FinNL, FinRE, FinFE, and FinNSP, using BERT with an additional classification layer for all tasks.

6.2 Experiment 1: Comparison of Pre-trained Model Architectures

For the two models in the general domain, GPT2-base and T5-base, their pre-training corpora, hyperparameters, and training volume are all the same, but their average scores differ significantly, with T5-base significantly outperforming GPT2-base, as shown in Table 4. This difference is mainly due to the differences in the architectures, parameter sizes, and pre-training methods of the T5 and GPT models. This performance confirms the correctness of our choice of the T5 model.

6.3 Experiment 2: Effectiveness of Domain Pre-training

As shown in Table 4, the comparison between the FinT5-base model and the T5-base model indicates that the FinT5-base model pre-trained on FinCorpus significantly outperforms the T5-base model with the same parameter size, demonstrating the effectiveness of domain pre-training and the effectiveness of FinCorpus.

6.4 Experiment 3: Superiority Compared to Existing Models in the domain

As shown in Table 4, in the four language understanding tasks evaluated with FinBERT and Mengzi-BERT-base-fin, FinT5-base significantly outperformed both models, demonstrating the superiority of FinT5 over existing models in the domain.

6.5 Experiment 4: Effectiveness of KETM

As shown in Table 4, by comparing FinT5-base-ke with FinT5-base, it can be seen that the knowledge-enhanced text modeling method significantly improves the model's performance on tasks such as relation extraction and news summarization, without significantly compromising the performance on other tasks, thus proving the effectiveness of the KETM method.

6.6 Experiment 5: Effectiveness of parameter scaling up

As shown in Table 4, the performance comparison between FinT5-base and FinT5-large models indicates that the FinT5-large model with one billion parameters performs significantly better than the FinT5-base model, demonstrating the effectiveness of parameter scaling up.

7 Conclusion

In this article, we introduced three new contributions to the domain of NLP in the context of Chinese finance. We created the largest open-source corpus for this domain, called FinCorpus, which contains a diverse collection of around 300GB of text from four sources. Our FinT5 model is the largest pre-trained language model for the Chinese financial domain, with one billion parameters. To enhance our pre-training method, we developed a unique knowledge-based approach called KETM,

which was effective. We also created a benchmark to evaluate the understanding and generation capabilities of language models, called CFLEB. We believe domain benchmarks should prioritize practicality to better reflect how improvements in language models in academia can benefit the real world. Our future work includes expanding FinCorpus and FinT5 and exploring multilingual and multimodal applications.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Biendata. 2019. Ccks 2019 extraction of public company announcement information.

Biendata. 2020a. Ccks 2020: Cross-class few-shot transfer event extraction for financial domain.

Biendata. 2020b. Ccks 2020: Evaluation of automated construction techniques for financial knowledge graph based on ontology.

Biendata. 2021. Ccks 2021: Event relation extraction for financial texts (part ii) - extraction of causal relationships between events.

Biendata. 2022a. Ccks2022: Evaluation of nl2sql for financial domain.

Biendata. 2022b. Ccks2022: Few-shot event extraction for financial domain.

Datafountain. 2019. Discovery of new entities in internet finance.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. Duee-fin: A large-scale dataset for document-level event extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 172–183. Springer.

Panpan Hou, Mengchao Zhang, Zhibing Fu, and Yu Li. 2020. Finbert. https://github.com/valuesimplex/FinBERT. GitHub repository, commit: ec1b14b96de9bdb5217abba1d197428cf00ddaa6.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. 2019. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*.

Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. 2019. Chinese relation extraction with multi-grained information and external linguistic knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4377–4386, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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. *arXiv preprint arXiv:1907.11692*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019a. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019b. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Alisa Smirnova and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, 51(5):1–35.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Tianchi. 2018. The dataset for extracting announcement information of a-share listed companies.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cndbpedia: A never-ending chinese knowledge extraction system. In *Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part II*, pages 428–438. Springer.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Jian Yang, Gang Xiao, Yulong Shen, Wei Jiang, Xinyu Hu, Ying Zhang, and Jinghui Peng. 2021. A survey of knowledge enhanced pre-trained models. *arXiv preprint arXiv:2110.00269*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.