FinBART: A Pre-trained Seq2seq Language Model for Chinese Financial Tasks

Hongyuan Dong[†], **Wanxiang Che**[*,†], **Xiaoyu He**[‡], **Guidong Zheng**[‡], **Junjie Wen**[‡]
[†]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
[‡]China Merchants Bank AILAB
{hydong, car}@ir.hit.edu.cn
{hexiaoyu42, zhengguidong, wenjunjieee}@cmbchina.com

Abstract

Pretrained language models are making a more profound impact on our lives than ever before. They exhibit promising performance on a variety of general domain Natural Language Processing (NLP) tasks. However, few work focuses on Chinese financial NLP tasks, which comprise a significant portion of social communication. To this end, we propose FinBART, a pretrained seq2seq language model for Chinese financial communication tasks. Experiments show that FinBART outperforms baseline models on a series of downstream tasks including text classification, sequence labeling and text generation. We further pretrain the model on customer service corpora, and results show that our model outperforms baseline models and achieves promising performance on various real world customer service text mining tasks.

1 Introduction

From making investment decisions to managing personal finances, financial text data is crucial in helping individuals and organizations stay up-to-date with the latest financial trends. With the development of information technology, the volume of available financial information looms so large that the use of machine learning methods to facilitate financial text processing becomes increasingly important. However, financial text data differs from general domain data as it contains a large number of technical terms and financial concepts, which require abundant expert knowledge to analyze and annotate. As a result, the application of powerful deep learning technology in financial text processing tasks is limited.

In this work, we apply the fast-developing pretraining techniques to Chinese financial domain. Although a number of works attempt to adapt transformer-based [1] pretrained language models to financial text [2, 3, 4], few studies focus on the application on Chinese financial text data. What's worse, these works adopt transformer encoder architecture, leading to limited ability in modeling the dependency between future tokens and poor performance in text generation tasks. To this end, we propose FinBART, which is an encoder-decoder transformer language model pretrained on Chinese financial corpora. FinBART extracts bi-directional semantic information with its encoder, and generates text with the decoder part. In this way, FinBART is capable of tackling both understanding and generation tasks in Chinese financial domain.

In real world scenarios of the financial industry, customer service is mainly conducted through text interaction. Therefore, there is a great demand for text analysis tools. Customer service corpora differ from financial ones for their sample text is highly colloquial. Although FinBART learns professional knowledge from the financial field, it cannot handle customer service tasks as well as financial ones without adaptation. To this end, we further pretrain FinBART model with customer service corpora, which are comprised of customers' queries about financial services. Further pretraining on customer service corpora adapts FinBART to text interaction scenarios, enabling FinBART to capture accurate semantic information of customer queries.

In summary, the main contributions of this work are listed as follows:

- We pretrain a transformer encoder-decoder language model named FinBART on Chinese financial corpora to infuse professional knowledge into the model. Experiments show that FinBART

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

outperforms baseline models on a variety of Chinese financial tasks, validating the necessity of domain-specific adaption.

- We futher pretrain FinBART on customer service corpora with a newly proposed pretraining objective. Experiments show that further pretrained FinBART-CS achieves higher performance on a series of customer service text mining tasks, facilitating text information filtering and processing in the customer service field.

2 Related Works

Unsupervised pretraining technique is regarded as one of the most significant breakthroughs in recent Natural Language Processing (NLP) research. Taking advantage of unsupervised pretraining methods, Pretrained Language Models (PLMs) achieve remarkable performance on a variety of natural language understanding and generation tasks. Decoder-only language models, represented by GPT [5, 6], possess a remarkable capability for text generation, but lack language understanding ability because they model semantic information in a causual way. Encoder-only language models like BERT [7], ELECTRA [8] and RoBERTa [9] are proposed to introduce bidirenctional semantic information to obtain more representative word embeddings for downstream tasks. However, Encoder-only models cannot model the dependency between `[MASK]` tokens and therefore suffer from limited generation capability. Encoder-decoder transformer models are more versatile because they extract bidirenctional semantic information with the encoder part and generate text smoothly with their decoders. Representative encoder-decoder models are BART [10] and T5 [11], which adopt different pretraining objectives to gain general language modeling ability.

Although it is not far from trivial to adapt language models pretrained on general corpus to domain-specific downstream tasks, their performance may be limited because of the lack of professional knowledge. This problem looms larger for domains involving large amounts of technical terms and concepts like finance. To this end, researchers propose to pretrain language models on domain-specific corpus to endow models with professional skills for downstream tasks. BioBERT [12] collects large biomedical corpora and pretrains a transformer encoder model to produce contextualized word representation for biomedical text. SciBERT [13] trains a BERT model for scientific NLP tasks with scientific publication corpora. In financial industry, researchers also seek to train domain-specifc language models to facilitate financial text meterials processing. FinBERT [2] designs a series of pretraining objectives for English financial text to train the model more effectively. Mengzi-BERT-fin [14] further trains Chinese BERT model to adapt to Chinese financial text. However, these works are mainly designed for Engligsh natural lnaguage understanding tasks and lack the ability to tackle Chinese NLP tasks involving generation. Our work fills this gap by pretraining a seq2seq transformer-based language model on Chinese financial corpora.

3 Methods

In this section, we introduce the pretraining procudure of the proposed FinBART and FinBART-CS. We choose transformer encoder-decoder architecture as the backbone, reconciling language understanding and generation ability. We use financial corpora and customer service corpora for model pretraining, and then finetune FinBART and FinBART-CS in their domain-specific downstream tasks, respectively. The overall data collecting, model pretraining and downstream task adapting pipeline is shown in Figure 1.

3.1 Copora

3.1.1 Chinese Financial Copora

We purchase a number of Chinese financial corpus data from Datayes, which is a financial technology data provider company. We also crawl a large amount of financial news from news portals and financial websites such as EastMoney[1], Ji Wei Net[2], and CNStocks Net[3]. In total, we collect an amount of 20

[1]https://www.eastmoney.com/
[2]https://m.laoyaoba.com/
[3]https://www.cnstock.com/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

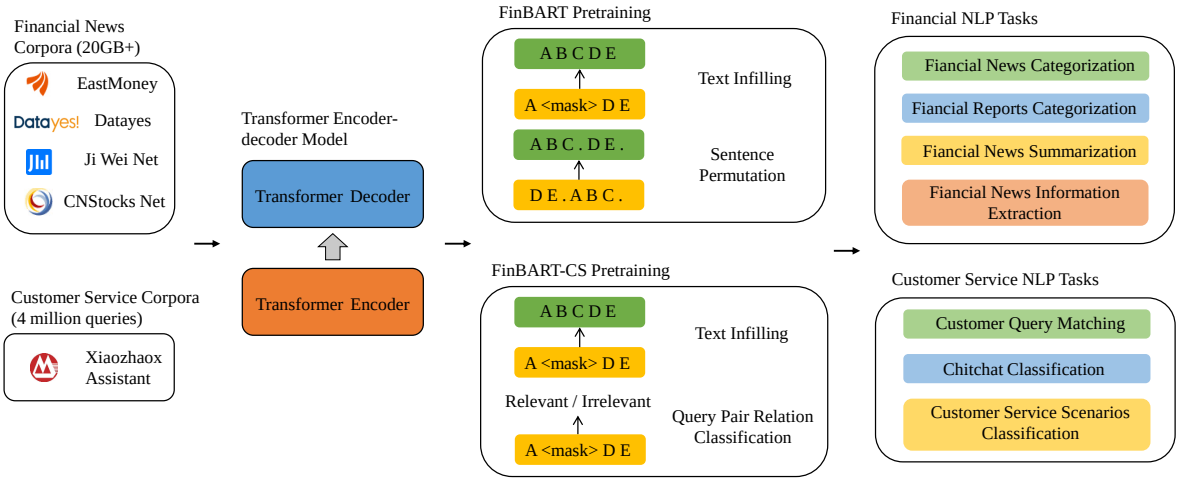

Figure 1: The overall pipeline of data collecting, FinBART and FinBART-CS model pretraining and downstream tasks adaptation.

GB Chinese financail news data, including approximately 12 million pieces of news and 1 billion tokens. These news are published between October 2020 to April 2022.

We further filter the financial new corpus with CCNet [15] toolkit, which extracts high quality datasets from web crawl data. To be specific, we use the LM filtering tool provided in CCNet to compute the perplexity of each financial news sample using a Chinese language model. News samples which score over 12,000 perplexity are filtered out. We also discard news samples which consist of more than 52% digits and puctuations, which are highly probable to be structured data and not expected to appear in the pretraining data.

3.1.2 Customer Service Query Copora

In customer service scenarios, the text content is highly colloquial and the language style and form are distinct from written financial news. To tackle real world customer service text mining tasks, we collect a large amount of customer service query data to further pretrain the model. We collect and mannually compile customer queries, and annotate each query with a primary question. Each query is semantic equivalent to its primary question. In this way, the customer service queries can be clustered into several semantic equivalent question groups, and each cluster is represented with a single primary question. We obtain approximately 4,038,304 customer service queries in total.

3.2 FinBART pretraining

We adopt BART [10] architecture as the model's backbone. BART is an encoder-decoder transformer model. It captures bidirenctional semantic information via self-attention mechanism with its encoder. Global context information is injected to the contextualized word embeddings produced by the encoder, which are then referred to with cross-attention during decoding.

As shown in Figure 2, we choose the combination of text infilling and sentence permutation as the pretraining objectives of our model, which is proven to be the most effective pretraining strategy for BART model [10]. For text infilling objective, we conduct whole word masking on Chinese financial documents to raise the difficulty, preventing the model from degrading to predict the masked word with only its neighbour words. We use Jieba, an open-source Chinese word segmentation toolkit, to segment financial documents into whole Chinese words. We mask a total of thirty percent of tokens and replace the contiguous whole word spans as single `[MASK]` tokens. For sentence permutation objective, we split financial documents into sentences, and shuffle the order of the sentences randomly. The model is trained to recover the sentence order on the decoder side, and is therefore driven to model the semantic information of the whole document.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

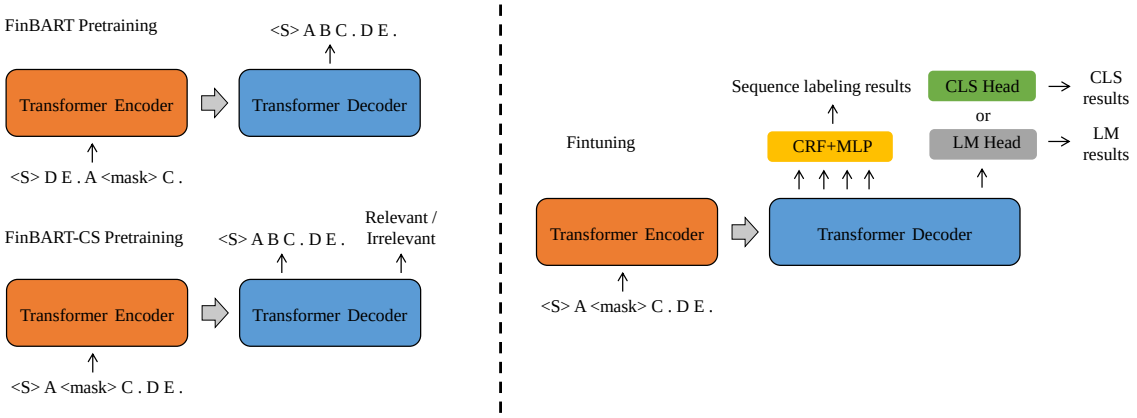

Figure 2: A detailed illustration of the pretraining (left) and finetuning procedure (right) of FinBART and FinBART-CS. "CLS" stands for classfication, and "LM" stands for language modeling. "CRF+MLP" is the abbreviation of conditional random field and multi-layer perceptron.

3.3 FinBART-CS pretraining

Customer service requires a large amount of text interaction, leading to an urgent need of automatic text mining tools. To this end, we futher pretrain FinBART with customer service corpora to facilitate domain-specific text analysis.

Pretraining data construction. We reorganize the customer service corpus into query pairs. To be specifc, we denote a query sample as $d = (\mathbf{x}_d, d^p)$, where $\mathbf{x}_d$ stands for the text content of the query and d^p is its primary question. Each query is of the same meaning as its primary question, and only one query is marked as primary question in its cluster. We denote a cluster with d^p as its primary question as follows:

$$\mathcal{C}_{d^p} = \{d_i | d_i^p = d^p\}. \tag{1}$$

For each query in the corpus $\mathcal{D}$, we sample five other queries pointing to the same primary question randomly from $\mathcal{C}_{d^p}$. These samples are paired with d to form five query pairs of equivalent semantic meaning. We denote the set of paired semantic equivalent query pairs as $\mathcal{D}_d^{eq}$. Similarly, we sample five queries from other clusters to form query pairs of two unrelated queries. We represent the unrelated query pairs as $\mathcal{D}_d^{dif}$. When pairing queries, we set the order of the two queries randomly. In this way, we obtain ten query pairs for each single query in the corpus. We aggeragate each query sample d's corresponding $\mathcal{D}_d^{eq}$ and $\mathcal{D}_d^{dif}$ to form the final pretraining dataset:

$$\mathcal{D}_{paired} = \bigcup_{d \in \mathcal{D}} \mathcal{D}_d^{eq} \cup \bigcup_{d \in \mathcal{D}} \mathcal{D}_d^{dif}. \tag{2}$$

Pretraining objectives. To utilize the knowledge in the customer service corpus effectively, we add a supervised training objective, which is query pair relation classification, in addition to BART's pretraining objective. The pretraining objectives of FinBART-CS is illustrated in Figure 2 (left). We extract the semantic representation of the given query pairs with the FinBART-CS model. Each query pair is represented as the hidden states of the last token extracted from the final decoder layer. An classification head, which is contructed with 2-layer neural network, is then used to map the representation to a binomial distribution indicating whether the pair of queries are of the same meaning. Model parameters are optimized with cross entropy loss between the predicted binomial distribution and ground truth label. We also conduct whole word masking for the text infilling objective, but discard sentence permutation objective because customer service queries are relatively short and not as coherent as written language. As a result, introducing sentence permutation objective cannot not promote the model to capture overall

semantic information, but instead confuses the model and leads to worse quality of semantic representations produced by the encoder model.

4 Experimetns

In this section, we introduce the experimental settings and resources used in the pre-training and downstream task adaptation stages of FinBART and FinBART-CS.

4.1 FinBART

4.1.1 Pretraining Settings

The proposed FinBART is a 12-layer transformer encoder-decoder model with 768-dimensional inner representation. The encoder and decoder of FinBART both consist of 6-layer transformer blocks while conducting different attention mechanism. FinBART adopts WordPiece [16] tokenizer to segment Chinese text.

We train two versions of FinBART, which are training from scratch and continual training. For training from scratch, we initialize model parameters randomly and optimize the parameters with pretraining loss computed on Chinese financial corpora. We set learning rate as 7e-4 with a batch size of 2048 and weight decay of 0.01. We use Adam optmizer for model pretraining with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, and warm up the pretraining procedure for the first 10,000 steps. After warming up, the learning rate decays linearly for 60,000 steps to a minimum learning rate of 1e-6. For continual training, we start from bart-base-chinese [17] checkpoint, and futher train the model for 800,000 steps. The learining rate is set as 2e-5, which warms up for the first 10,000 steps and decays for 500,000 steps liniearly.

We implement FinBART pretraining pipeline with Megatron-LM [18] framework. We use data parallel strategy on 8 NVIDIA V100 GPUs to accelerate FinBART pretraining. Mixed precision training is also adopted in the pretraining procedure, where model parameters and garadients are stored in FP16 precision, while the accumulated gradients and parameter optimization are computed in FP32 precision. A gradient upscaling operation is also adopted to alleviate the parameter underflow problem.

4.1.2 Downstream Tasks

We collect four downstream tasks to evaluate FinBART's capabilities in Chinese financial NLP tasks. For each task, we finetune the model on the training set with a learning rate of 1e-5 and a batch size of 8, and evaluate the model's performance on validation set with a fixed interval. When the validation performance does not increase for more than 20 times of valiadation, we stop finetuning and test the model's final performance on the test set.

Financial news categorization. We leave out a small proportion of the Chinese financial news pretraining corpus to evaluate FinBART's natural language understanding ability. These financial news are collected from news portals and financial websites from October 2020 to April 2022. Financial news categorization dataset is a 29-class text classification task, which contains 4,292 training samples, 537 validation samples and 537 testing samples belonging to different industires. The average length of sample texts is 430 words. During finetuning, we extract sentence embedding as the hidden states of the last token, and implement a 2-layer classfication head to transform the representation to classification results. We use categorization accuracy as the evaluation metric.

Financial reports categorization. We collect a financial reports categorization dataset to evaluate model's performance on more professional financial documents. Financial reports categorization dataset is a 41-class text classification task, with a training set of 3,330 samples, a validation set of 416 sampels and a test set of 417 samples. The average length of sample texts is 430 words. We process financial reports categorization in the same way as financial news categorization task and use categorization accuracy as the evaluation metric.

Financial news summarization. We also leave out a small proportion of financial news corpus to test model's text generation ability. For each financial news document, we use the article body as input text, and select its abstraction as the target summarization. Financial news summarization dataset contains

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Model	News Cat Accuracy	Reports Cat Accuracy	News Sum BLEU	DUEE-fin Slot F1
MENGZI-BERT-BASE-FIN [14]	79.14	76.74	–	**81.01**
BART-BASE-CHINESE [17]	78.21	76.98	0.6141	76.27
FINBART (SCRATCH)	**81.19**	79.62	0.6087	76.08
FINBART (CONTINUAL)	80.07	**80.10**	**0.6206**	76.44

Table 1: Raw experiment results of FinBART on four Chinese financial downstream tasks. Bold numbers indicate the best result on the task.

Model	News Cat	Reports Cat	News Sum	DUEE-fin	Average
MENGZI-BERT-BASE-FIN [14]	+1.19	−0.31	–	**+6.21**	–
BART-BASE-CHINESE [17]	0.00	0.00	0.00	0.00	0.00
FINBART (SCRATCH)	**+3.81**	+3.43	−0.88	−0.25	+1.53
FINBART (CONTINUAL)	+2.38	**+4.05**	**+1.06**	+0.22	**+1.93**

Table 2: Normalized relative performance improvement of FinBART over BART-base-chinese baseline on four Chinese financial downstream tasks. The results are given in percentage. Bold numbers indicate the best result on the task.

39,694 trainin samples, 4,962 validation sampels and 4,962 test samples. The average length of the article body is 367 words, and reference summarizations' average length is 82 words. We adopt greedy decoding strategy to generate summarizations, and restrict the vocab to tokens appeared in the original news text. The results are evaluated with BLEU [19] score between the generated text and the reference.

Duee-fin. Duee-fin is a publicly available financial event extraction dataset [20], where each sample contains a Chinese financial news, corresponding event type and argument roles. In total, there are 13 event types and 61 slot labels appearing in Duee-fin dataset. Duee-fin contains 9,425 training samples, 1,518 validation samples and 273,876 test samples. Because we cannot get access to labels of the test set, we sample training, validation and test set randomly from the union of original training and validation sets according to a ratio of 8:1:1. We use the official demo pipeline to preprocess the event, extracting the trigger sentence and its corresponding sequence labels in BIO format. We use the sequence labels to evaluate FinBART's capability in token-level financial NLP tasks. As shown in Figure 2 (right), we implement a linear transformation to map each token's last layer representation to a distribution over sequence labels. A Conditional Random Field (CRF) is then used to model the dependency between sequence labels. We use macro F1 score over all slot types as the evaluation metric.

4.1.3 Main Results

We report the raw experimental results for each task in Table 1. Since each task adopts different metrics, the overall improvement of the proposed models is hard to be quantified. To address this issue, we follow UL2 [21] to use the *normalized relative gain with respect to baselines* as the overall metric. We use BART-base-chinese as the baseline and compute the relative performance improvement of other models. The normalized results are listed in Table 2.

As listed in Table 1, compared with BART-base-chinese [17] trained on generic Chinese corpora and Mengzi-bert-base-fin [14] trained on financial corpora, FinBART not only tackles both financial language understanding and generation tasks, but also achieves better performance on these datasets. For News categorization dataset, the best model FinBART trained from scratch gains 81.19 classification accuracy, which is 3.81% higher than BART-base-chinese baseline and 2.59% higher than Mengzi-bert-base-fin baseline. For Reports categorization dataset, continually trained FinBART achieves the best 80.10 classfication accuracy, outperforming BART-base-chinese with 3.12 accuracy score (4.05%↑) and leading Mengzi-bert-base-fin with 4.38% higher accuracy. For financial news summarization dataset,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Training strategy	Bsz	Lr	# Steps	News Cat	Reports Cat	News Sum	DUEE-fin
–	–	–	–	78.21	76.98	0.6141	76.27
SCRATCH	256	1e-4	1e6	**81.75**	77.70	0.5729	**78.51**
	2048	7e-4	2e5	81.19	79.62	0.6087	76.08
CONTINUAL	256	2e-5	8e5	80.07	80.10	**0.6206**	76.44
	2048	1e-4	8e4	78.21	**80.34**	0.6024	77.94

Table 3: The influence of hyper parameters to FinBART's performance on downstream Chinese financial NLP tasks. The first line shows the performance of BART-base-chinese baseline. Bold number is the best result on the task.

Training strategy	Bsz	Lr	# Steps	News Cat	Reports Cat	News Sum	DUEE-fin	Average
–	–	–	–	78.21	76.98	0.6141	76.27	0.00
SCRATCH	256	1e-4	1e6	**+4.53%**	+0.94%	−6.71%	**+2.94%**	+0.42%
	2048	7e-4	2e5	+3.81%	+3.43%	−0.88%	−0.25%	+1.53%
CONTINUAL	256	2e-5	8e5	+2.38%	+4.05%	**+1.06%**	+0.22%	**+1.93%**
	2048	1e-4	8e4	0.00%	**+4.36%**	−1.91%	+2.19%	1.16%

Table 4: Relative performance gain of FinBART with different hyper parameter settings over BART-base-chinese baseline. The first line shows the performance of BART-base-chinese baseline. Bold number is the best result on the task.

continually trained FinBART achieves the best performance with 0.6206 BLEU score, surpassing BART-base-chinese with a 1.06% margin. Mengzi-bert-base-fin cannot tackle text generation tasks because of its transformer encoder architecture, so its performance on news summarization task is not listed in the table. For Duee-fin task, Mengzi-bert-base-fin achieves the highest 81.01 slot F1 score. We attribute its advantage to its bi-directional language modeling mechanism. Among language models with encoder-decoder architecture, continually trained FinBART performs the best with 76.44 slot F1 score, which is 0.22% higher than BART-base-chinese baseline.

To facilitate comparison, we reorganize the experiment results into relative performance improvement over BART-base-chinese baseline. The reorganized results are shown in Table 2. We show the average relative performance gain in the last column to compare the overall performance of each model. Generally speaking, FinBART models trained on Chinese financial corpora achieve better results than BART-base-chinese baseline, indicating the necessity of adaptation training on domain-specific corpora. FinBART continually trained from BART-base-chinese checkpoint gains the most relative performance improvement, leading BART-base-chinese baseline with a 1.93% higher relative performance averagely on the four Chinese financial NLP tasks. FinBART trained from scratch on Chinese financial corpus outperforms BART-base-chinese baseline with 1.53% higher overall performance, but fails to win BART-base-chinese on 2 out of 4 tasks. We ascribe its unsatisfactory performance to the small size of the pretraining corpus, which results into limited general language modeling ability. Continually trained FinBART contains both general knowledge and financial domain-specific knowledge, and therefore outperforms BART-base-chinese baseline with a larger margin on all four downstream tasks.

4.1.4 The Impact of Hyper Parameters

In Section 4.1.1, we introduce the different sets of hyper parameters when training FinBART from scratch and continually training from BART-base-chinese checkpoint. To understand how hyper parameters influence the model performance, we train FinBART under different hyperparameter settings. We set varying pretraining batch size, learning rate and the number of pretraining steps for pretraining. For FinBART trained from scratch, we train the model with 1e-4 and 7e-4 learning rate respectively, and set batch size and optimization steps accordingly for fair comparison. For FinBART continually trained from BART-base-chinese checkpoint, we use smaller learning rates as suggested for BERT domain adaptation

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

training code. The results are shown in Table 3 and Table 4.

Overall, FinBART models continually trained from BART-base-chinese checkpoint achieve better results than those trained from scratch. We attribute the leading performance of continually trained FinBART models to the large corpora used in the pretraing process. The two stage pretraining (general domain pretraining & financial domain continual pretraining) procudure injects both general knowledge and financial domain-specific knowledge into the model. Therefore, continually trained FinBART models show impressive ability in Chinese financial NLP tasks.

For FinBART models trained from scratch, we find that the combination of large batch size and large learning rate obtains better model performance, which is consistent with the conclusion of RoBERTa [9]. For continually trained FinBART, we set smaller learning rates for domain adaptation training. In this circumstance, the model trained with small batch size and more training steps achieves the best performance.

4.2 FinBART-CS

4.2.1 Pretraining Settings

FinBART-CS shares the same architecture with FinBART, which is a 12-layer encoder-decoder transformer model with 768-dimensional inner representation. We train FinBART-CS with additional customer service corpora from FinBART checkpoint. We set the learning rate as 2e-5 and train the model for 1,000,000 steps with a 10,000 steps warming up and 500,000 steps linear learning rate decay. The minimum learning rate is 1e-6. We use 256 batch size for FinBART-CS pretraining, and optimizes model parameters with Adam optimizer. FinBART-CS's optimization involves two pretraining objectives, which are text infilling and query pair relation classification. We give equal weights to the two objective losses, and optimize FinBART-CS parameters and classification head parameters jointly.

4.2.2 Downstream Tasks

We collect four downstream tasks to evaluate FinBART-CS's capabilities in customer service NLP tasks. For each task, we finetune the model on the training set with a learning rate of 1e-5, and monitor model performance with the validation set to avoid overfitting. We set the early stopping threshold as 20.

BQ Corpus. BQ Corpus is a publicly available customer service query matching dataset [22]. Each sample consists of a pair of customer service queries and a label indicating whether the two queries have the same semantic meaning. We concatenate the query pair and feed it to the model, converting it to a binary text classification task. BQ Corpus dataset contains 88,000 training sampels, 11,000 validation samples and 11,000 testing sampels. The average length of the query pairs is 34. We set the finetuning batch size as 8 and use classification accuracy as the evaluation metric.

Query match. Query match dataset's formulation is akin to that of BQ Corpus. Each sample consists of a pair of customer queries and a binary label indicating whether the query pair shares the same semantic meaning. Query match dataset consists of a 2,967,316-sample training set, a 370,914-sample validation set and a 370,915-sample test set. The average sample text length is 36. We set the finetuning batch size as 32 and use classification accuracy as the evaluation metric.

Chitchat classification. Chitchat classification dataset is a binary text classfication task. Sample text is obtained from real world Xiaozhaox smart assistant customer service scenario. Each sample contains a piece of customer query text and a label indicating whether the query is asking for services or just chitchatting. Chitchat classification dataset contains 102,459 samples, and we split it into training, validation and testing parts randomly according to a ratio of 8:1:1. The average length of the sample text is 14. We set the finetuning batch size as 8 and use classification accuracy as the evaluation metric.

Xiaozhaox categorization. Xiaozhaox categorization is a 1,046-class text categorization task. Sample text is obtained from real world Xiaozhaox smart assistant customer service scenario. Note that although a small proportion of samples of Xiaozhaox categorization dataset are contained in FinBART-CS pretraining corpora, categorization label information is not introduced during pretraining. Each sample belongs to a certain class of 1,046 catrogories. Xiaozhaox categorization dataset contains 1,837,722

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Model	BQ Copus Accuracy	Query Match Accuracy	Chitchat Cls Accuracy	Xiaozhaox Cls Accuracy
MENGZI-BERT-BASE-FIN [14]	92.17	88.38	99.29	65.40
BART-BASE-CHINESE [17]	88.82	87.58	99.07	63.38
FINBART (SCRATCH)	90.75	87.15	99.08	62.08
FINBART (CONTINUAL)	91.65	87.43	99.07	63.63
FINBART-CS (CONTINUAL)	**93.32**	**90.97**	**99.36**	**67.94**

Table 5: Raw experiment results of FinBART-CS on four Chinese financial customer service downstream tasks. Bold numbers indicate the best result on the task.

Model	BQ Copus	Query Match	Chitchat Cls	Xiaozhaox Cls	Average
MENGZI-BERT-BASE-FIN [14]	+3.77%	+0.91%	+0.22%	+3.19%	+2.02%
BART-BASE-CHINESE [17]	0.00	0.00	0.00	0.00	0.00
FINBART (SCRATCH)	+2.17%	−0.49%	+0.01%	−2.05%	−0.09%
FINBART (CONTINUAL)	+3.19%	−0.17%	0.00%	+0.39%	+0.85%
FINBART-CS (CONTINUAL)	**+5.07%**	**+3.87%**	**+0.29%**	**+7.19%**	**+4.11%**

Table 6: Normalized relative performance improvement of FinBART-CS over BART-base-chinese baseline on four Chinese financial downstream tasks. The results are given in percentage. Bold numbers indicate the best result on the task.

training samples, 30,000 valiadation samples and 6,029 testing samples. The average length of the text samples is 13. We set the finetuning batch size as 32 and use classification accuracy as the evaluation metric.

4.2.3 Results

We report the raw experimental results for each task in Table 5. We also reorganize the experiment results into *normalized relative gain with respect to baselines* to compare the model's overall performance. The normalized model performance scores are listed in Tabel 6.

As shown in Table 5 and Table 6, FinBART-CS continually trained on customer service corpus achieves the highest accuracy across all four Chinese financial customer service NLP tasks, and outperforms BART-base-chinese baseline with 5.07%, 3.87%, 0.29% and 7.19% accuracies, respectively. Averagely, continually pretraining on customer service corpus leads to a 4.11% overall performance improvement. Meanwhile, we have following observations based on Table 5 and Table 6:

- FinBART-CS is the only model outperforming Mengzi-bert-base-fin. Mengzi-bert-base-fin adopts transformer encoder-only architecture, which is regarded more suitable for natural language understanding tasks. None of the listed encoder-decoder model gains higher scores on the four datasets except FinBART-CS, indicating the superiority of FinBART-CS in customer service NLP tasks. Compared to Mengzi-bert-base-fin, FinBART-CS learns to understand customer service text data via continual pretraining, making up the lack of bi-directional semantic information.

- Models trained with financial corpora achieve better performance on financial customer service tasks. Continual pertraining on financial corpora injects financial domain-specifc knowledge into the model, facilliating the understanding and processing of financial customer service text. That being said, customer service text data is highly colloquial, and its distribution is different from that of written financial news data. As a result, FinBART trained on both general corpora and financial corpora outperforms FinBART trained from scratch with only financial corpora, which may overfit to the distribution of financial document data.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Model	BQ Copus	Query Match	Chitchat Cls	Xiaozhaox Cls
BART-BASE-CHINESE	88.82	87.58	99.07	63.38
FINBART-CS	**93.32**	**90.97**	**99.36**	**67.94**
FINBART-CS (ABLATION)	91.65	87.43	99.07	63.63

Table 7: FinBART-CS and classfication loss ablated FinBART-CS's performance on four financial customer service NLP tasks. Bold numbers indicate the best result on the task.

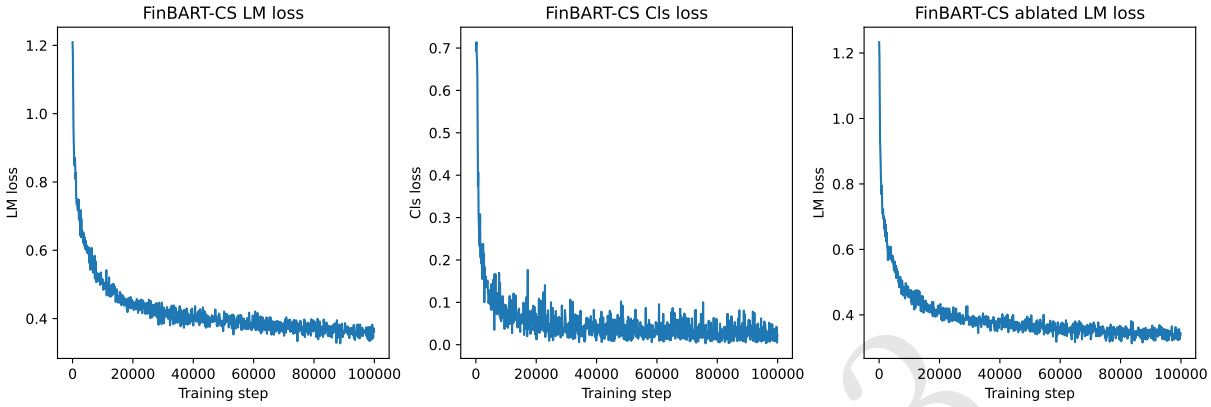

Figure 3: The loss curves of FinBART-CS's LM loss (left), FinBART-CS's Cls loss (mid) and FinBART-CS's ablated LM loss (right). "LM" stands for language modeling, and "Cls" stands for classfication. We curate the first 1e5 steps of the pretraining procedure for clarity.

4.2.4 Ablation Study

To validate the effectiveness of the proposed pretraining objective for FinBART-CS, we conduct abaltion study for the query pair relation classification objective. We train the ablated FinBART-CS with the same set of parameters as FinBART-CS with only query pair relation classification loss removed from the pretraining procedure. Table 7 shows the performance of FinBART and FinBART-CS on four Chinese financial customer service NLP tasks. When query pair relation classification objective is removed, the ablated FinBART-CS suffers from a large performance drop. FinBART-CS outperforms the ablated model on all four downstream tasks, validating the effectiveness of query pair relation classification pretraining objective.

We also investigate how the newly introduced classification loss influences FinBART-CS's pretraining procedure. As shown in Figure 3, the LM (Language Modeling) loss of FinBART-CS converges smoothly regardless of whether the classification loss is introduced. The query pair relation classification loss also converges fast and smoothly during the pretraining process. Supervised signals are injected to the model without hindering the model from learning general language knowledge. The loss curves show the compatibility of LM loss and classification loss, validating the effectiveness of our proposed query pair relation classification pretraining objective.

5 Conclusions

In this work, we propose FinBART and FinBART-CS, which are Chinese encoder-decoder language models pretrained on Chinese financial corpora and customer service corpora, respectively. The proposed models fill the blank of Chinese seq2seq language model for financial and customer service NLP tasks. Experiments on a series of Chinese financial NLP tasks and customer service NLP tasks indicate the promising performance of the proposed models. We also conduct analysis experiments to show the rationality of pretraining hyper parameter selection and the effectiveness of the proposed pretraining objective for FinBART-CS.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, 2021.

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

[4] Vinicio DeSola, Kevin Hanna, and Pri Nonis. Finbert: pre-trained model on sec filings for financial natural language tasks. *University of California*, 2019.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019. Association for Computational Linguistics.

[8] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692, 2019.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[13] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. 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 3615–3620, 2019.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

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

[15] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

[16] Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144, 2016.

[17] Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*, 2021.

[18] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[20] Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. Duee-fin: A large-scale dataset for document-level event extraction. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, pages 172–183. Springer, 2022.

[21] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022.

[22] Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.