Gradual Further Pre-training Architecture for Economics/Finance Domain Adaptation of Language Model

Hiroki Sakaji
Department of Systems Innovation, School of Engineering
The University of Tokyo
Tokyo, Japan
sakaji@sys.t.u-tokyo.ac.jp

Masahiro Suzuki
Department of Systems Innovation, School of Engineering
The University of Tokyo
Tokyo, Japan

Kiyoshi Izumi
Department of Systems Innovation, School of Engineering
The University of Tokyo
Tokyo, Japan

Hiroyuki Mitsugi
Data-driven Science Dept.
Frontier Technologies Research Center.
Daiwa Institute of Research Ltd.
Tokyo, Japan

***Abstract*—We propose a new pre-trained architecture for economics/finance domain adaptation of language models in this paper. Pre-trained language models have become commonplace in a wide range of language processing applications. Because they learn from generic documents such as Wikipedia, many pre-trained language models are not fully adapted to the domain. As a result, there are two approaches: one is to develop a domain-specific pre-trained language model, and the other is to adapt the model learned on general documents to the domain through further pre-training with domain texts. However, no definitive better method has been discovered, and each project is working on it in different ways. As a result, this study focuses on the Japanese language's economics/financial domains and investigates how pre-trained language models can be better adapted to domain-specific tasks.**

***Index Terms*—Pre-trained Language Model, Further Training, Domain Adaptation**

I. Introduction

Pre-trained language models have become commonplace in a wide range of language processing applications. It is used for both general documents and domain texts such as economic/financial texts. However, because they learn from generic documents such as Wikipedia, many pre-trained language models are not fully adapted to the domain. As a result, there are two approaches: one is to develop a domain-specific pre-trained language model, and the other is to adapt the model learned on general documents to the domain through further pre-training with domain texts. However, no clear superior solution has been discovered, and each project is working on it in different ways.

In the economics/finance domain, language processing is becoming increasingly crucial. Chatbots and financial document summaries, for example, have been employed practically. However, new economics/finance domain-specific tasks are currently being developed, such as sentiment analysis in economics/finance. One tough element is that economic/financial feeling differs from general sentiment. For example, restructuring is a bad term in general, but it is a good term in economics/finance. As a result, improved language processing approaches in economics/finance domain-specific activities are required.

Therefore, this study focuses on the Japanese language's economics/financial domains and investigates how pre-trained language models might be better fitted to domain-specific tasks. We believe that by having the models learn progressively increasingly specialized texts, beginning with generic texts, a good domain-specific model can be constructed. To validate this assumption, multiple models are built and tested using experiments on three tasks and four datasets.

Our main contributions are summarized here.

- For building the pre-trained language model, we proposed a domain adaptation architecture.
- To demonstrate the utility of the suggested strategy, we built multiple models and ran experiments on three tasks with four additional datasets.

II. Related Works

BERT [1] contributed significantly to the pre-learning model. The basic BERT architecture consists of 12 encoder layers of transformer. With a larger corpus and increased computational complexity, XLNet [2] based on Transformer-XL [3] outperformed BERT. ALBERT [4] enhanced BERT by modifying its parameters and tasks. By sharing the parameters of each layer and decomposing the embedding matrix, ALBERT decreases the number of parameters. RoBERTa [5] improved on BERT's tuning approach. It removed NSP and performed dynamic masking with large minibatches and also improved the input format. DistilBERT [6] was proposed based on the concept of knowledge distillation [7].

Gururangan et al. [8] explored domain adaptation of pre-trained models employing four domain texts in their study on domain adaptation of pre-trained learning models (biomedical and computer science publications, news, and reviews). Peng et al. [9] investigated the utility of domain adaptation for numerous financial activities using BERT and FinBERT. Furthermore, they concentrated on the vocabulary of pre-trained models and investigated if changes in vocabulary affected performance. In contrast, we suggested a revolutionary additional pre-trained architecture for language model adaptation in the economics/finance.

III. Methodology

In this section, we will discuss how to build domain-specific pre-trained language models as well as the model for classifying aspect-based sentiments using pre-trained language models that have been built. One of the three objectives is aspect-based sentiment analysis which is discussed here because the model is more characteristic than the other two. Other models for the other two tasks only conduct fine-tuning using the output of the [CLS] tag.

A. Pre-training

BERT is pre-trained through multitasking fill-in-the-blank questions (MLM) and judgment on the connectivity of two sentences (NSP). The model and training parameters are the same as in the original BERT work [1]; we utilize the BERT base with a batch size of 256, a learning rate of 1e-4, several training steps of 1M, and an optimizer of AdamW (Adam with weight decay). In contrast to the original paper, we train the model using a sequence lenght of 512 for all steps. We build pre-trained models using the repository[1], which is based on the PyTorch-based framework[2], [3].

B. Further Training Architecture

We postulate that a good domain-specific model may be developed by having the models learn progressively more specialized texts, beginning with generic texts. As a result, we propose a new training architecture as seen in Figure 1. In Figure 1, the model is trained by Wikipedia as the first phase and then by the general financial documents as the second step. Finally, as the third stage, the model is further pre-trained using the economist reports. According to the original BERT implementation[4], we utilize 2e-5 as the learning rate for futher pre-training.

C. The model for aspect-based sentiment analysis

We present a methodology for classifying aspect-based sentiments. The model is shown in Figure 2. In this study, we input a sentence and a target for solving aspect-based sentiment analysis. We begin by generating a word sequence W from the input sentence. In this case, the word sequence

[1]https://github.com/retarfi/language-pretraining/tree/v1.0
[2]https://github.com/pytorch/pytorch
[3]https://github.com/huggingface/transformers
[4]https://github.com/google-research/bert

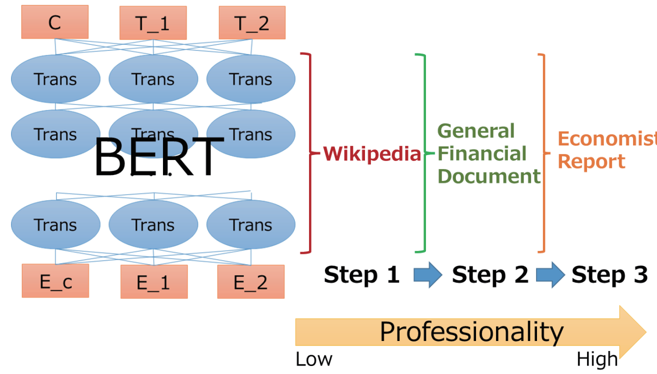

Fig. 1. Our further training architecture

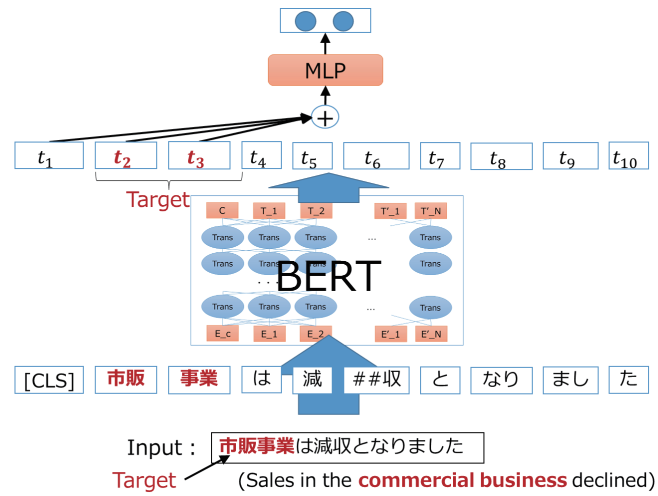

Fig. 2. The model for aspect-based sentiment analysis

W consists of N words; thus, $W = \{w_1, w_2, ..., w_N\}$. Then, we add [CLS] and [SEP] special tokens to W. Equation 1 calculates the vector T_W of each word using this input and the BERT model $BERT$ trained on Japanese Wikipedia entries.

$$T_W = BERT(W). \tag{1}$$

where $T_W = \{t_1, t_2, ..., t_N\}$ and the dimension l of the final BERT output is 768, and $t_i \in \mathbb{R}^l$. Then, we choose t that corresponds to the [CLS] tag and the words in the input target, and we define the set of these t as M. In Figure 2, t_1, t_2, t_3 are selected as M.

$$y = W_M \sum_{m \in M} m + b_M, \tag{2}$$

where y is an output layer, W_M is a weight matrix, and b_M is a bias vector. We employ cross-entropy as a loss function and adam as an optimizer.

IV. Experiment

We assess our built pre-trained language models using three tasks and four datasets. We use Wikipedia (as of June 1, 2021), general financial documents, and economist reports to build pre-trained language models. As general financial documents, we obtained Japanese financial results and Japanese securities reports. The financial results are derived from TDnet[5] and cover the period from October 9, 2012, to December 31, 2020. The securities reports were obtained from EDINET[6] and covered the period from February 8, 2018 to December 31, 2020. Daiwa Institute of Research Ltd provided the economist reports, which include 53,595 documents from 2003 to 2020. We evaluate nine models to confirm the performance of our further training architecture. There are models available that do not rely on economist reports[7].

Wiki
: is a BERT model pre-trained using Wikipedia.

Wiki+Fin
: is a BERT model pre-trained using Wikipedia and general financial documents.

Wiki+Eco
: is a BERT model pre-trained using Wikipedia and economist reports.

Wiki+Fin+Eco
: is a BERT model pre-trained using Wikipedia, general financial documents and economist reports.

Wiki $\rightarrow$ Fin
: is a BERT model pre-trained using Wikipedia and further pre-trained using general financial documents.

Wiki $\rightarrow$ Eco
: is a BERT model pre-trained using Wikipedia and further pre-trained using economist reports.

Wiki $\rightarrow$ Fin+Eco
: is a BERT model pre-trained using Wikipedia and further pre-trained using general financial documents and economist reports.

Wiki+Fin $\rightarrow$ Eco
: is a BERT model pre-trained using Wikipedia and general financial documents, and further pre-trained using economist reports.

Wiki $\rightarrow$ Fin $\rightarrow$ Eco
: is a model further pre-training Wiki $\rightarrow$ Fin using economist reports.

We perform fivefold cross-validation with fixed test data in the experiment. The training data is divided into five parts one of which is used as validation data to train the model. The test data is then used to run five tests, and the average of the evaluation indices is calculated.

In this experiment, we perform three tasks with four datasets. One of them is the detection of causality. We use two datasets in this task: a Japanese financial statement summary known as "Kessan Tanshin" and Nikkei newspaper articles, which are Japanese newspaper articles. Other tasks include aspect-based sentiment analysis and economic sentiment analysis. A chABSA dataset[8] is used for aspect-based sentiment analysis. We use a Japanese economy watchers survey[9] for economic sentiment analysis which provides a timely and accurate overview of regional economic trends. The following are the specifics of three tasks.

Causality detection
: seeks to detect causality in a financial sentence input. This task is described in detail in our previous paper [10]. The datasets we employed are Kessan Tanshin and Nikkei newspaper articles. We collected the Kessan Tanshin data from Japanese company websites and generated the dataset by randomly selecting 30 files. To create the experimental causality dataset, we tagged the data, which was done by two natural language processing researchers (and they are also individual investors). Each researcher labeled tags for the 30 files, and the results of the two researchers were combined. As a result, the data consists of 2,891 sentences 394 of which are causality sentences, and we divided the data into 2,313 training sentences (303 causality sentences) and 578 test sentences (91 causality sentences). We randomly selected 2,045 sentences from Nikkei newspaper articles from 1995 to 2005; the sentences were then tagged by five annotators. We obtained the sentences that had been labeled as causality sentences by three or more annotators. As a result, we obtained 898 causality sentences, which we divided the sentences into 1,632 training sentences (709 causality sentences) and 413 test sentences (189 causality sentences).

Aspect-based sentiment analysis (ABSA)
: The chABSA dataset contains 7,723 sentiments, 4,334 positives, 3,131 negatives, and 258 neutrals. We use only positives and negatives in this study because the number of neutrals is insufficient. We divided the data into 5,973 training samples (3,456 positives and 2,517 negatives) and 1,492 test samples (868 positives and 624 negatives).

Economic sentiment analysis (ESA)
: The economy watchers survey includes documents and their economic sentiments. The economic sentiment are given in 5 levels. Therefore, in this task, we classify input documents into five classes. We acquired the economy watchers survey from 2010 to 2017. And, we selected 30,000 data from the acqruired economy watchers survey, randomly. Then, we split the data into 25,000 and 5,000 test data.

A. *Results and Discussion*

The experiment results are shown in Table I. Moreover, an example of the result of causality detection by "Wiki $\rightarrow$ Fin $\rightarrow$ Eco" and "Wiki" is shown in Table II. In Table I, Kessan means Kessan Tanshin, Newspaper means Nikkei newspaper, Economy Watchers means economy watchers survey. Additionally, in the table, P means precision, R means recall, and each result represents the mean of macro values.

From Table I, the proposed pre-trained language model (Wiki $\rightarrow$ Fin $\rightarrow$ Eco) outperfomed other pre-trained language

[5] https://www.release.tdnet.info/inbs/I_main_00.html

[6] https://disclosure.edinet-fsa.go.jp/

[7] https://huggingface.co/izumi-lab

[8] https://github.com/chakki-works/chABSA-dataset

[9] https://www5.cao.go.jp/keizai3/watcher-e/index-e.html

TABLE I
EXPERIMENTAL RESULTS.

	Causality Detection						ABSA			ESA		
	Kessan			Newspaper			chABSA			Economy Watchers		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Wiki	0.878	0.854	0.864	0.870	0.867	0.866	0.870	0.867	0.868	0.454	0.438	0.426
Wiki+Fin	0.846	0.879	0.861	0.882	0.879	0.880	0.880	0.879	0.880	0.453	0.437	0.426
Wiki+Eco	0.870	0.895	0.882	0.882	0.881	0.881	0.879	0.882	0.880	0.485	0.421	0.411
Wiki+Fin+Eco	0.860	0.856	0.857	**0.886**	**0.885**	0.885	0.890	0.885	0.887	0.409	0.411	0.393
Wiki → Fin	0.866	0.869	0.867	0.851	0.842	0.842	0.871	0.871	0.871	0.381	0.372	0.350
Wiki → Eco	0.854	0.853	0.853	0.877	0.871	0.873	0.866	0.864	0.865	0.439	0.425	0.411
Wiki → Fin+Eco	0.848	0.886	0.864	0.882	0.882	0.882	0.887	0.886	0.886	0.478	0.435	0.420
Wiki+Fin → Eco	0.861	0.872	0.864	**0.886**	0.884	0.885	0.879	0.881	0.879	**0.504**	0.436	0.427
Wiki → Fin → Eco	**0.894**	**0.889**	**0.891**	**0.886**	**0.885**	**0.886**	**0.891**	**0.887**	**0.888**	0.500	**0.440**	**0.429**

TABLE II
AN EXAMPLE OF A RESULT OF CAUSALITY DETECTION.

	Correct	Output	Sentence
Wiki → Fin → Eco	1	1	昨年発生した東日本大震災の復興需要等を背景に緩やかな回復を続けております
Wiki	1	0	(A gradual recovery continues against the backdrop of reconstruction demand after the Great East Japan Earthquake that occurred last year)

models on all tasks. From the results, we confirm that our assumption is appropriate for economics/finance domain adaptation of pre-trained language models. In addition, as shown in Table 2, we discovered some cases where the proposed model detected causality that other models did not. Other tasks produce comparable results. Additionally, we compared the model that further trained the economist reports with other models to examine the words that tended to be included in the correct answers. Consequently, we found that 需要(demand), 情報(information), 経済(economics) were included. Therefore, we consider that further training using economist reposts empowered weights of domain words.

V. Conclusion

We proposed a new pre-trained architecture for economics/finance domain adaptation of language models in this paper. We believe that learning specialized texts progressively, beginning with general texts, could result in a good domain-specific model. Based on this concept, we tested several pre-trained language models using three tasks and four datasets. The results of the experiments confirmed that our assumption is useful for domain adaptation of language models.

We intend to build our pre-trained model for other languages such as English in the future. In addition, we will attempt to tackle other tasks such as summarization using pre-trained language models that we have built.

Acknowledgments

This work was supported by Daiwa Securities Group. This work was supported in part by JSPS KAKENHI Grant Number JP21K12010, and the JST-Mirai Program Grant Number JPMJMI20B1, Japan.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[2] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[3] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2978–2988.

[4] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[7] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[8] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8342–8360.

[9] B. Peng, E. Chersoni, Y.-Y. Hsu, and C.-R. Huang, "Is domain adaptation worth your investment? comparing bert and finbert on financial tasks," in *Proceedings of the Third Workshop on Economics and Natural Language Processing*, 2021, pp. 37–44.

[10] H. Sakaji, R. Murono, H. Sakai, J. Bennett, and K. Izumi, "Discovery of rare causal knowledge from financial statement summaries," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–7.