Constructing a Korean Named Entity Recognition Dataset for the Financial Domain using Active Learning

Dong-Ho Jeong[a,O], **Min-Kang Heo**[a], **Hyung-Chul Kim**[b], **Sang-Won Park**[a]
DeepNatural Inc[a], Kookmin Bank[b]
dongho@deepnatural.ai, minkang@deepnatural.ai, yhdosu@kbfg.com, anson@deepnatural.ai

Abstract

The performance of deep learning models depends on the quality and quantity of data. Data construction, however, is time-consuming and costly. In addition, when expert domain data are constructed, the availability of experts is limited. In such cases, active learning can efficiently increase the performance of the learning models with minimal data construction. Although various datasets have been constructed using active learning techniques, vigorous studies on the construction of Korean data on expert domains are yet to be conducted. In this study, a corpus for named entity recognition was constructed for the financial domain using the active learning technique. The contributions of the study are as follows. (1) It was verified that the active learning technique could effectively construct the named entity recognition corpus for the financial domain, and (2) a named entity recognizer for the financial domain was developed. Data of 8,043 sentences were constructed using the proposed method, and the performance of the named entity recognizer reached 80.84%. Moreover, the proposed method reduced data construction costs by 12–25%.

1 Introduction

Rapid advancements in the field of artificial intelligence have contributed to their increased use in various fields. In the field of natural language processing, deep learning models have demonstrated a higher performance than humans in benchmark tests such as GLUE (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2018). Such deep learning models require expert knowledge and time, in addition to the cost involved in computing power for model training and data construction. Therefore, it is difficult to apply deep learning models in industries readily. In particular, it is not easy to reduce the costs of learning data construction when compared to that of hardware because it requires human effort. In addition, specific fields, such as the financial domain, requires workers with expert domain knowledge and, therefore, requires more time and is more expensive. Various studies have been conducted in areas such as unsupervised learning (Taghipour and Tou Ng., 2015) and crowdsourcing (OOmen and Aroyo., 2011) to reduce the data construction costs of expert domains. The number of data constructed through such methods is larger than the number of annotations created by experts, but the quality of the constructed data is relatively low.

Active learning is a technique used by experts to construct data, with the goal of achieving optimum performance with fewer data. The active learning technique determines which data are to be learned first by interacting with users when learning data are limited. The goal of the interaction is to identify data that can be used to improve the performance of the learning model efficiently. During the early learning stages, the model is trained with a small number of learning data, and it submits queries requesting additional data that can efficiently improve its performance, thereby reducing development costs. The active learning technique is known to be effective in

Proceedings of the 17th International Conference on Natural Language Processing, pages 208–212
Patna, India, December 18 - 21, 2020.

various natural language processing fields such as information extraction, Named Entity Recognition(NER), and text classification (Settles., 2009). Various studies have been conducted on active learning for general and expert domains in other countries, but vigorous studies for constructing expert domain data are yet to be conducted in South Korea.

This study investigates active learning-based data construction for a NER system in the financial domain. The contributions of this study are as follows.

(1) It was experimentally verified that the active learning technique could effectively construct a NER corpus for the financial domain. Particularly, a cost reduction of 12.5–25% was achieved when the data of 3,000–3,500 sentences were constructed.

(2) A named entity recognizer for the financial domain with a performance of 80.84% was consequently developed.

The remainder of this paper is organized as follows. Section 2 presents a review of existing studies on active learning from Korea and other countries and an examination of the research directions in active learning. Section 3 describes the system diagram for constructing the NER corpus for the financial domain. The corpus and results are evaluated in Section 4, and the conclusions are presented in Section 5.

2 Related Research

2.1 Active Learning Technique

Passive learning represents general machine learning that learns data in sequence without a process of selecting learning data through queries. In contrast, active learning arbitrarily selects and learns data. Compared to passive learning, more learning data selection schemes exist in active learning (Settles., 2009). Uncertainty sampling is the simplest way to select the initial learning data. It is an intuitive way of assuming that if the model prediction for specific data is uncertain, information on the corresponding data is insufficient for selecting the data as the priority target data to be learned. Uncertainty sampling schemes include least confidence (LC), margin sampling, and entropy sampling according to uncertainty-measuring criteria. The most intuitive LC sampling scheme, which uses data with the lowest measured probability (e.g., softmax) first, is used in this study to develop the trained model. In other words, it first selects data that are not easily resolved by the current model as the learning data.

2.2 Research Trends in Korea and Other Countries

Studies on active learning are being conducted in various natural language processing fields, including NER (Shen et al., 2017), and in various domains (Li et al., 2012), including the financial domain (Smailović et al., 2014), in other countries (Settles., 2009). In South Korea, the active learning technique has been applied to areas such as sentence classification (Kim et al., 2012) and NER (Yoon and Oh., 2015) using person, location, and organization (PLO) representation of the named entities. However, it has not been vigorously applied to domains that require expert knowledge, such as the financial domain. In this study, active learning was applied to the construction of a NER corpus for the financial domain to recognize 40 named entities (financial institutions, broadcasting stations, economy-related institutions, and financial products such as funds and derivatives) in the finance-related domain.

3 Construction of a Named Entity Corpus for the Financial Domain using Active Learning

In this study, the NER corpus was constructed by applying active learning to the financial domain. NER identifies and classifies PLO entities in a given sentence. In the financial domain, this is an annotation task that distinguishes financial and general institutions and requires expert knowledge for tagging named entities of financial products (e.g., funds) and financial theories and phenomena. However, there are practical limitations such as the requirement for experts who perform the annotation tasks and the costs of data construction. Therefore, in this study, we investigate whether the performance of a model can be improved by minimizing the annotation tasks using active learning to handle such problems. Figure 1 presents the overall diagram of the named entity corpus construction task for the financial domain using active learning.

The task sequence is as follows. (1) First, 500 arbitrary data, which are to be annotated, are selected from the financial domain raw corpus. (2)

A named entity annotation task is performed on the corresponding data by an annotation platform. (3) The created data are then used to train the model. (4) The model is used to predict the raw corpus and select the annotation target data belonging to a specific condition. The LC scheme discussed in Section 2.1 was applied to select the annotation target data in this study. Steps (1) - (4) are repeated until the performance of the model converges. The model selects data that are expected to contribute the most to model performance by learning the sentences that are not initially predicted by the model. In other words, as data of a specific size or larger tend to contribute less to performance improvement, the model performance can be optimized with minimum data. As a result, a better performance can be obtained using our method than using random sampling, even with the construction of fewer data at a reduced cost.

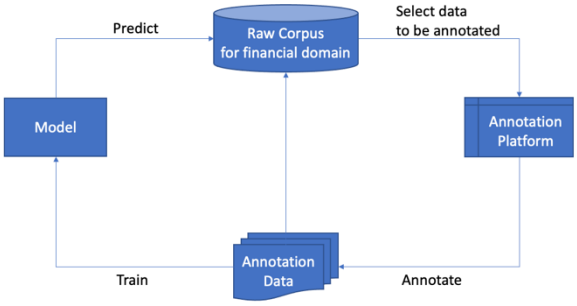

Figure 1: Schematic diagram for constructing the named entity corpus for the financial domain using active learning.

4 Experiment and Results

4.1 Data

Financial domain text was collected from the web to verify the effectiveness of the active learning model. [1] The text was collected by crawling sentences, including finance-related keywords. Of the 9,043 sentences collected, named entities in 1,043 sentences were tagged first and used as evaluation data, and the remaining 8,000 sentences were treated as the raw financial domain corpus. The tagging task was performed with 40 named entity tag sets using the annotation platform [2] shown in Figure 2.

4.2 Experiment Method

The initial annotation target data were randomly selected from the raw financial domain corpus. The selected data were tagged by the annotation platform shown in Figure 2. A morpheme analyzer provided by KB was used to separate the annotated sentences into morpheme units for use as learning data. In this study, a named entity recognizer was implemented by fine-tuning the multilingual-BERT model using the original method in (Devlin et al., 2015).

The experimental method is as follows: The model was trained using the active learning technique that samples the target data using LC and random sampling methods. Thereafter, the model was trained and evaluated by increasing the number of learning sentences by 500 to examine the performance improvement trend in the model following each iteration. The final model was generated in eight iterations, with the data increasing from 500 to 4,000 sentences. In addition, each iteration was repeated 10 times to verify the objectivity of the experiment. The hyperparameters used in the experiment are listed in Table 1.

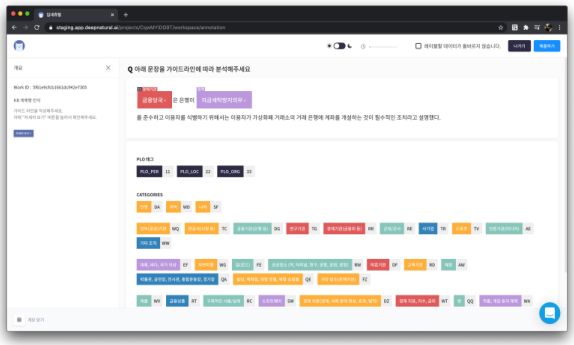

Figure 2: Processing tools of the annotation platform.

Hyperparameter	Value
Maximum sentence length	256
Batch size	6
Learning rate	3e-5
Optimizer	Adam
Epoch	10

Table 1: Font Hyperparameter information.

[1] The dataset was collected in cooperation with KB.

[2] http://app.deepnatural.ai

4.3 Experiment Results

The F1 evaluation results of the active learning model trained using LC and random sampling are presented in Figure 3. When the number of learning sentences is small (500–1,500 sentences), there are no significant differences in the model performance. It appears that at least 1,500 sentences are required to obtain meaningful information to affect the model performance. When the number of learning sentences exceeds 2,000, the model trained using the LC method performs better than that trained using random sampling. With 4,000 or more sentences, the LC and random sampling models tend to exhibit similar performances. Specifically, the F1 of random sampling is 79%, which is similar to that when the LC model is trained with 3,000–3,500 sentences. This shows that the cost can be reduced by 12.5-25% by using the active learning technique.

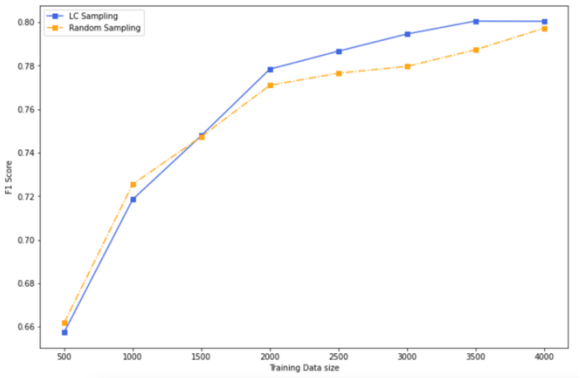

Figure 3: Model performance evaluation trend using different sampling techniques

Information on the tag set ratios of 500 sentences selected in each of the eight iterations for generating the annotation target data is presented in Table 2. Only the first four iterations are listed in Table 2 for brevity.

Step	Top 5 tag ratios (%) in LC sampling	Low 5 tag ratios (%) in random sampling
#1	AMOUNT (25)	AMOUNT (20)
	PRICE (20)	DATETIME (17)
	DATETIME (15)	PRICE (15)
	FINANCIAL_INSTITUTION (10)	FINANCIAL_INSTITUTION (15)
	PERSON (4)	PERSON (5)
#2	FINANCIAL_INSTITUTION (15)	AMOUNT (24)
	COMPANY_AND_BRAND (8)	PRICE (17)
	GOVERNMENT (7)	DATETIME (17)
	FINANCIAL_PRODUCT (7)	FINANCIAL_INSTITUTION (10)
	VALUE (6)	VALUE (6)
#3	FINANCIAL_INSTITUTION (15)	AMOUNT (23)
	DATETIME (12)	PRICE (16)
	AMOUNT (11)	DATETIME (14)
	AREA (6)	FINANCIAL_INSTITUTION (13)
	GOVERNMENT (5)	PERSON (5)
#4	PRICE (21)	AMOUNT (28)
	FINANCIAL_INSTITUTION (14)	PRICE (19)
	DATETIME (11)	DATETIME (13)
	AMOUNT (8)	FINANCIAL_INSTITUTION (9)
	FINANCIAL_PRODUCT (6)	VALUE (5)

Table 2: Top 5 tag set ratios (%) in the first four step.

The tag sets ‘AMOUNT,’ ‘PRICE,’ ‘DATETIME,’ and ‘FINANCIAL_INSTITUTION’ were mainly selected by random sampling, indicating that these tag sets frequently appear in the collected raw corpus. The tag set distribution for the entire annotated sentences also reveals that the tags ‘DATETIME’, ‘AMOUNT,’ ‘FINANCIAL_INSTITUTION’ and ‘PRICE’ appear frequently. However, in LC sampling, various tag sets appear as the upper tag set regardless of the tag set ratios in the entire dataset.

5 Conclusions

In this study, a NER corpus for the financial domain was constructed through active learning. The active learning technique could rapidly and efficiently improve the performance of the model. Tag sets that could not be easily analyzed by the model were constructed first, and it was found that more meaningful quality datasets were constructed through machine learning. From the 8,043 sentences constructed, a named entity recognizer with a F1 performance of 80.84% was developed for the financial domain. Further, it was experimentally verified that the active learning technique could provide a cost reduction of 12.5–25%.

Acknowledgments

This study was carried out as part of the 2019 Startup Growth Technology Development Project (TIPS Program, No. S2816383) supported by the Ministry of SMEs and Startups in cooperation with KB Kookmin Bank.

References

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Kaveh Taghipour, and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In: *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*. p. 314-323.

Johan Oomen, and Lora Aroyo. 2011. Crowdsourcing in the cultural heritage domain: opportunities and challenges. In: *Proceedings of the 5th International Conference on Communities and Technologies*. p. 138-149.

Burr Settles. 2009. *Active learning literature survey: Computer sciences technical report 1648*. University of Wisconsin-Madison.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Lianghao Li, Xiaoming Jin, Sinno Jialin Pan, and JianTao Sun. 2012. Multi-domain active learning for text classification. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. p. 1086-1094.

Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. 2014. Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, 285, 181-203.

Je-wook Kim, Han-joon Kim, and Sang-goo Lee. 2012. Construction method of active learning-based learning document sets for a Bayesian document classification system. *Journal of KIISE: Software and Application*, 29(11· 12), 966-978.

Bo-hyeon Yoon and Hyo-jeong Oh. 2015. Development of a tool for semi-automatically constructing named entities in advance using active learning. *Journal of KACE*, 18(6), 81-88.

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*.