# Optimizing Domain Specificity of Transformer-based Language Models for Extractive Summarization of Financial News Articles in Korean

**Huije Lee** **Wonsuk Yang** **ChaeHun Park** **Hoyun Song**
**Eugene Jang** **Jong C. Park***
School of Computing
Korea Advanced Institute of Science and Technology
{jae4258,derrick0511,ddehun,hysong,eugenej,park}@nlp.kaist.ac.kr

## Abstract

Frequent usage of complex expressions with numbers and of the terms that require domain knowledge makes it more difficult to comprehend and summarize financial news articles than that of other daily news articles. We present a transformer-based model for the automatic summarization of the financial news articles in Korean and address related issues, and in particular analyze the interplay between the domain of the dataset used for pre-training and that for fine-tuning. We find that the summarization model performs much better when the two coincide, even when they are different from that of the target task, which is the financial domain in our work.

## 1 Introduction

It has been widely acknowledged that some pre-trained language models can be effective for tasks with little training data, as pre-trained models are often able to achieve moderate (and sometimes satisfactory) performance, even without a large dataset that had been commonly employed to train neural models before the introduction of such pre-trained language models (Radford et al., 2019; Zhang et al., 2020). It has also been acknowledged that such pre-trained language models are effective for low resource languages (Conneau and Lample, 2019). Also, language-agnostic tokenization methods such as Byte-Pair Encoding (Sennrich et al., 2016) and SentencePiece (Kudo and Richardson, 2018) have been introduced, making it possible to fine-tune language models without language-specific tokenizers.

Many of the recently introduced language models are based on transformers (Vaswani et al., 2017). Among them, pre-trained language models gained much attention due to their advanced performance on various tasks. Specialized language models have also been introduced to domains such as biomedicine (Lee et al., 2020), clinics (Huang et al., 2019), and finance (Liu et al., 2020).

When we use such a pre-trained language model for a domain-specific task, the language model can be expected to perform better if the model is pre-trained specifically for the domain at hand. However, it has been recently noted that it is non-trivial to determine whether systems that use a pre-trained model trained on a domain-specific corpus may outperform those that use a model pre-trained on a general domain corpus (Lee et al., 2020; Gururangan et al., 2020)

In this work, we address the summarization task for financial news articles in Korean. We follow extractive summarization rather than abstractive summarization, as the latter can mistakenly use some numbers or terminology that do not appear in the original document (Kryscinski et al., 2019). In contrast, extractive summarization only selects sentences (or phrases) from the original document, without such 'hallucinations' (Koay et al., 2020).

We analyze the effectiveness of a model based on the domain specificities of the datasets that are used for pre-training and for fine-tuning, as shown in Figure 1. We use (1) Korean Wikipedia, (2) financial news articles in Korean, and (3) non-financial

* Corresponding author

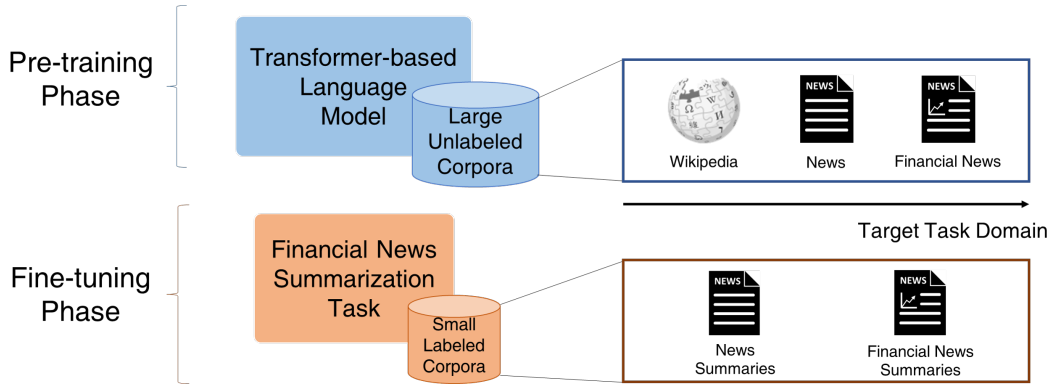


Figure 1: Schematic of the two-phase training framework for domain specific models

news articles in Korean, all for the pretraining of the model, and use the 'Modu' news summarization dataset published by the Korean government, which includes human-generated gold standards for summarization. Our analysis shows that the performance of the models on the task of summarization of financial news articles improves even when the synchronized domain is not finance. To the best of our knowledge, this is the first work on the impact of the domain similarity (or difference) between the datasets for pre-training and fine-tuning, in particular on the performance gain (or loss) of the summarization model for financial news articles in Korean.

Our contributions in this work are as follows.

- We present domain-specific models for automatic summarization of financial news articles in Korean.
- We present an in-depth analysis of the interplay between the domains of the datasets that are used for pre-training and fine-tuning of a summarization model.
- We present a case study for the summaries automatically generated by our models (and a few strong baselines), comparing them with the human-generated gold standards.

## 2 Related Work

### 2.1 Pre-trained Language Models

Gururangan et al. (2020) reported that domain-adaptive pre-training (DAPT), a secondary pre-training phase utilizing domain-specific texts, can increase performance. Additionally, they observed decreases in performance when datasets from other task domains were used in the adaptive pre-training phase.

Lee et al. (2020) proposed a pre-trained language model specialized for tasks in biomedicine. Heung et al. (2019) proposed a BERT model pre-trained on the clinical notes of patients for the task of 30-day hospital re-admission prediction. Liu et al. (2020) proposed a multi-task pre-training method on a financial corpus, a Wikipedia corpus, and a book corpus to improve performance on financial tasks. Beltagy et al. (2019) used a corpus of scientific publications for pre-training, and achieved high performance on core NLP tasks such as named entity recognition and dependency parsing.

Our work focuses on exploring the impact of domain differences for extractive summarization. We use a Transformer encoder-based architecture to analyze the performances of models by dividing the training into pre-training and fine-tuning phases.

### 2.2 Extractive Summarization

Rossiello et al. (2017) presented C-W2V, which selects important sentences close to the centroid embedding computed by the composition of word embeddings. Nallapati et al. (2016) proposed a two-layer recurrent neural network sequence classifier comprising a word-level layer and a sentence-level layer. Yao et al. (2018) and Narayan et al. (2018) proposed extractive summarization models that use reinforcement learning based on hierarchical network architectures. Dong et al. (2018) addressed extractive summarization as a contextual bandit prob-

lem and proposed a model that uses reinforcement learning on a sequence model with neural networks. Liu et al. (2019) presented BertSumExt, which identifies the sentences that play an important role in conveying the content of the document based on sentence embeddings ranked by BERT. Narayan et al. (2020) proposed HiBERT and ETCSum with a stepwise structured transformer that puts a sentence encoder and a documenter encoder together. Zhong et al. (2020) addressed extractive summarization through a summary level framework that evaluates a summary as a whole, rather than at sentence-level. Jeon et al. (2019) proposed a two-step hierarchical extractive summarization model utilizing the BERT model and bi-directional long short-term networks.

## 3 Domain-specific training in Two Phases

Our approach aims to analyze how domain similarity between the datasets influences the performance gain of text summarization, both on pre-training and fine-tuning. We use three different datasets for the pre-training and two other datasets for the fine-tuning. To quantitatively evaluate the domain difference among the pre-training datasets, we measure their vocabulary overlaps.

### 3.1 Corpora for Pre-training

For pre-training of the domain-specific language models, we put together three pre-training corpora: Korean Wikipedia articles, financial news articles, and non-financial news articles.

To collect news articles, we used an API provided by NAVER, a search engine[1]. Each news article was tagged with one of the seven categories: (1) World, (2) Society, (3) Politics, (4) Life/Culture, (5) IT/Science, (6) Opinion, and (7) Economy. We used news articles in the economy category to find financial news. We also used Korean Wikipedia articles for pre-training. Table 1 shows the statistics of the collected text data.

To clean up the text data, we deleted (1) the content outside of the article, such as the names of the publishers or the reporters, advertisements, or disclaimers, (2) duplicate paragraphs in an article due to an API error, and (3) characters that are neither numbers, English alphabets, Korean and/or Chinese alphabets, nor ASCII special characters.

[1]https://developers.naver.com/docs/search/news/

| Domain | # Tokens | # Sents | Size |
|---|---|---|---|
| Wikipedia | 61.9M | 4.4M | 0.6GB |
| News (¬F) | 3.3B | 19.1M | 29.9GB |
| News (F) | 3.2B | 17.7M | 29.3GB |

Table 1: Statistics for the pre-training corpora. *News (F)* and *News (¬F)* indicate the news articles in and out of the domain of finance, respectively.

| Domain | # Tokens | # Sents | # Articles |
|---|---|---|---|
| News (¬F) | 958K | 70.6K | 3,726 |
| News (F) | 194K | 13.9K | 619 |

Table 2: Statistics for fine-tuning datasets in the *Modu* news summarization dataset.

We measured the domain similarity between Korean Wikipedia articles and the news articles based on the overlaps among the 10,000 most frequently occurring words, excluding stopwords[2]. Specifically, we measured the similarity (1) between Korean Wikipedia articles and non-financial news articles, (2) between Korean Wikipedia articles and financial news articles, and (3) between non-financial news articles and financial news articles. We took the exact match of the token as a token overlap, and regarded two tokens that are different only in the prefix as different tokens. To measure domain similarity, we used the entirety of the Korean Wikipedia articles (0.63GB) and randomly selected the financial news articles and the non-financial news articles until the two sets are of the same size (0.63GB) as the Korean Wikipedia articles. Figure 2 shows the ratio of the token overlaps, where the overlap ratio between financial news and non-financial news is found higher than that of financial news and Wikipedia articles. This analysis illustrates that the corpus of non-financial news articles could be considered as a near-target domain corpus for summarization of financial news articles.

| Domain | Train | Valid | Test | Total |
|---|---|---|---|---|
| News (¬F) | 2,608 | 361 | 757 | 3,726 |
| News (F) | 432 | 72 | 115 | 619 |
| All | 3,040 | 433 | 872 | 4,345 |

Table 3: The size of fine-tuning datasets divided into training, validation, and test sets.

### 3.2 Corpora for Fine-tuning

We use the news summarization dataset included in the *Modu* dataset[3] for fine-tuning the summarization of financial news articles in Korean. The dataset consists of 4,389 news articles. For each article, the dataset provides a human-generated gold standard for extractive summarization, consisting of three sentences (selected from the original news article), taking into account their order as well so that the concatenation of the sentences can be presented as the gold standard for summarization. Among the 4,389 news articles, some articles were redacted or retracted, as they were not accessible via the Web. We excluded these 44 non-accessible articles. As a result, 4,345 news articles in total were used for our experiments, among which 619 news articles were assigned the finance category (F), as shown in Table 2. We split the news articles into train, valid, and test data with the ratio of 70%, 10%, and 20%, respectively. Table 3 shows the statistics of the dataset.

## 4 Method

We describe methods for building a summarization model in two phases: pre-training and fine-tuning.

### 4.1 Pre-training

**Masked Language Modeling (MLM)** randomly selects some tokens from the input sequence and replaces them with the masked token ($[MASK]$), where the model learns to restore the masked token into the original token. For the given input sequence $\bar{X} = \{x_1, x_2, ..., x_T\}$, the method creates a noised sequence, which contains the masked tokens, $\hat{X} = \{x_1, x_2, ..., [MASK]_i, x_{i+1}, ..., x_T\}$, where $i$ is in the range of $[1, T]$. The model learns to correctly predict $x_i$ from $\hat{X}$ through minimization of the cross-entropy loss. The loss function can be defined as in Equation 1, where $m_t$ is an integer that can be either 1 or 0, which indicates whether or not the $t$-th token $x_t$ is masked.

$$\mathcal{L}_{mlm} = \sum_{t=1}^{T} m_t \log p(x_t|\hat{X}; \theta) \quad (1)$$

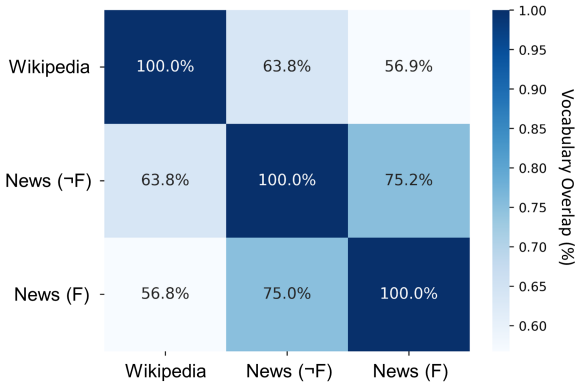


Figure 2: Vocabulary overlap among the three datasets. The vocabulary consists of top 10K most frequent words, excluding stopwords, sampled from 0.63GB of documents in each domain.

### 4.2 Fine-tuning

We share the view of Liu and Lapata (2019) and consider the extractive summarization task as a classification task, which classifies whether a given sentence in the original document should be part of the summarization result. We fine-tuned our model to minimize the binary cross-entropy loss for classification. For the ordering of the classified sentences, we sorted the sentences based on the likelihood (the value given by the output node) that represents the probability that the sentence is correctly classified as the target so that the more probable sentence comes earlier in the summarization results.

For a sentence, where $x_i$ is a token, we use MLM-style token representation to seamlessly fine-tune the pre-trained language model. Our model uses mean pooling for sentence representation: The average of the output representations of all the tokens within the input sentence is considered as the sen-

[2]https://gist.github.com/spikeekips/40eea22ef4a89f629abd87eed535ac6a

[3]https://corpus.korean.go.kr/

| Models | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Lead-3 | .48±0.00 | .25±0.00 | .38±0.00 |
| KoBERT | .58±.003 | .39±.004 | .50±.003 |
| BERT-Multilingual | .58±.007 | .40±.011 | .50±.010 |
| BERT-Multilingual-plus-finance | .57±.001 | .37±.001 | .49±.002 |
| MLM-zero-finance (ours) | .60±.002 | .43±.001 | .53±.001 |
| MLM-half-finance (ours) | .60±.002 | .44±.001 | .54±.003 |
| MLM-entirely-finance (ours) | **.62±.001** | **.46±.001** | **.56±.000** |

Table 4: Evaluation results of extractive models. The best scores in each metric are highlighted in **bold**.

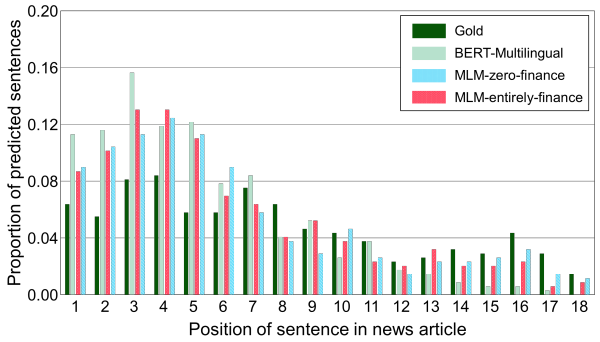


Figure 3: Proportion distribution of predicted sentence positions.

tence representation of the input sentence[4].

## 5 Experiment

We implement a few strong baseline models along with our models for automatic summarization and evaluate the models on 115 financial news articles[5] that are included in the *Modu* automatic news summarization dataset. The implementation details for the models are as follows.

### 5.1 Setting

#### 5.1.1 Pre-training

We use the following models as baselines in pre-training the summarization model.

**KoBERT**[6] is a transformer-encoder-based model that is trained on Korean Wikipedia articles. It has 12 transformer-encoder layers ($L = 12$) and a hidden dimension size of 768 ($H = 768$). The number of attention heads is 12 ($A = 12$), and the upper limit on the number of the tokens that can be in a sentence is 512. A drop-out rate of 0.1 is applied during the training. The model architecture is similar to that of *BERT-Base* (Devlin et al., 2019). The number of learnable parameters is 92M and the size of the dictionary is 8K, which are smaller than those of *BERT-Base*, 110M and 29K, respectively.

**BERT-Multilingual** (Devlin et al., 2019) is a transformer-encoder-based model that is trained on multi-lingual Wikipedia, including Korean.

We use the *bert-base-multilingual-cased*[7] among some variations of the pre-trained models. The architecture of the model is the same as that of the *BERT-Base*. The number of learnable parameters is 179M, and the number of the tokens in the dictionary is 119.5K, which are bigger than those of *BERT-Base* for English, 110M and 29K, respectively.

**BERT-Multilingual-plus-finance** is a model that uses financial news data to additionally pre-train on top of the pre-trained parameters of the BERT-Multilingual model described above. The dataset for the additional pre-training was the financial news articles that we gathered, and the size of the dataset is about 30GB, which we describe in detail in Section 3. The size of the batch was 32, which is the same as that used for pre-training BERT-Multilingual, and the learning rate was $5e$-5. We additionally pre-trained the model for 6 days, using two NVIDIA

[4]We also used sentence pooling (Liu and Lapata, 2019) via $[CLS]$ tokens at the beginning of each sentence for all the settings of the experiments. However, as the mean pooling showed better performance for almost all the settings, we only report the models with mean pooling.

[5]20% of all the financial news articles in the dataset, considered as the test set. The detailed information of the train/valid/test data can be found in Section 3.

[6]https://github.com/SKTBrain/KoBERT

[7]https://huggingface.co/bert-base-multilingual-cased

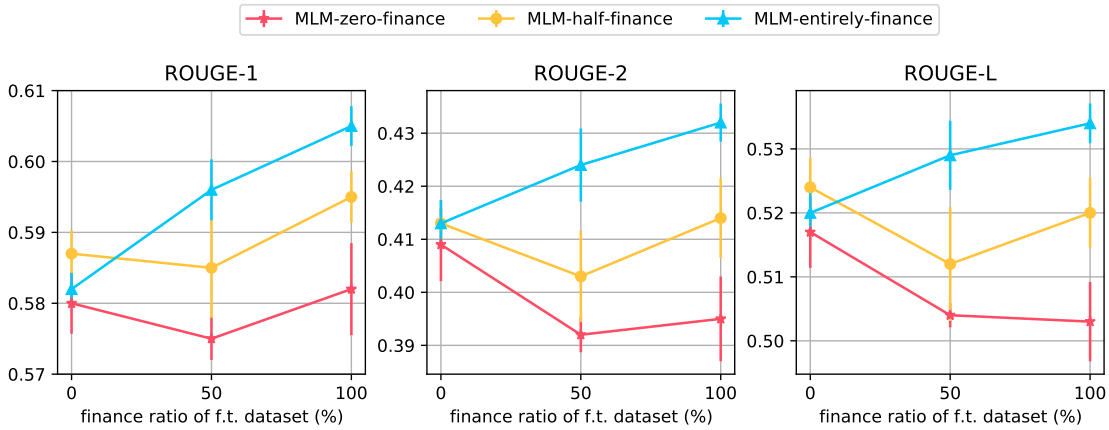


Figure 4: Evaluation results with varying ratios of financial news articles for training (both pre-training and fine-tuning). The X-axis in the plot indicates the ratio of financial news articles in fine-tuning. 'f.t.' refers to fine tuning.

V100 GPUs.

The details of our pre-trained language models for automatic summarization are as follows.

**MLM-zero-finance (Ours)** is a language model that is trained using our method based on Masked Language Modeling, which is explained in detail in Section 4. The non-financial news articles (29.9GB) were used for pre-training. For tokenization, we used the WordPiece tokenizer with the upper limit on the size of the dictionary set to 32K, the number of the merged tokens to 6K, and the minimum token frequency for the merger to 2. It has $L = 12$, $H = 768$, $A = 12$, and the upper limit on the number of tokens in an input sentence is 512. The size of the batch was 1K[8]. The learning rate was $5e$-5. We pre-trained the model for 6 days, using two NVIDIA V100 GPUs.

**MLM-half-finance (Ours)** is a language model that is pre-trained in the same manner as MLM-zero-finance except for the dataset on which it was trained. Half of the News ($\neg$F) collection and the other half of the News (F) collection were used for pre-training.

**MLM-entirely-finance (Ours)** is a language model that is pre-trained in the same manner as MLM-zero-finance except for the dataset. The financial news articles (29.3GB) were used for pre-training.

#### 5.1.2 Fine-tuning

We use Adam optimizer (Kingma and Ba, 2015) and linear learning rate decay, and the learning rate was $2e$-6. The batch size was 32.

### 5.2 Result

For the evaluation metric, we use the F1-score of ROUGE (Lin, 2004). ROUGE-1 measures the unigram overlap between the predicted and the gold result. ROUGE-2 measures the bigram overlap between the two, and ROUGE-L measures the longest common subsequence between the two. ROUGE-L can be interpreted as a fluency measure.

#### 5.2.1 Impact of Dataset Domain on Pre-training

Table 4 shows the performance of the models when they are fine-tuned on all of the available training data[9] (3,040 news articles) and when they are evaluated against the human-generated gold standards for the 115 financial news articles (the test data introduced in Section 3.2).

MLM-entirely-finance showed the best performance for ROUGE-1, ROUGE-2, and ROUGE-L. MLM-half-finance showed the next best perfor-

[8]Liu et al. (2019) and You et al. (2020) reported that a large size batch can help the performance of a pre-trained language model.

[9]We believe that this setting of using all of the available training datasets is most similar to that in the development scenario for a real-world application of summarization to financial news articles.

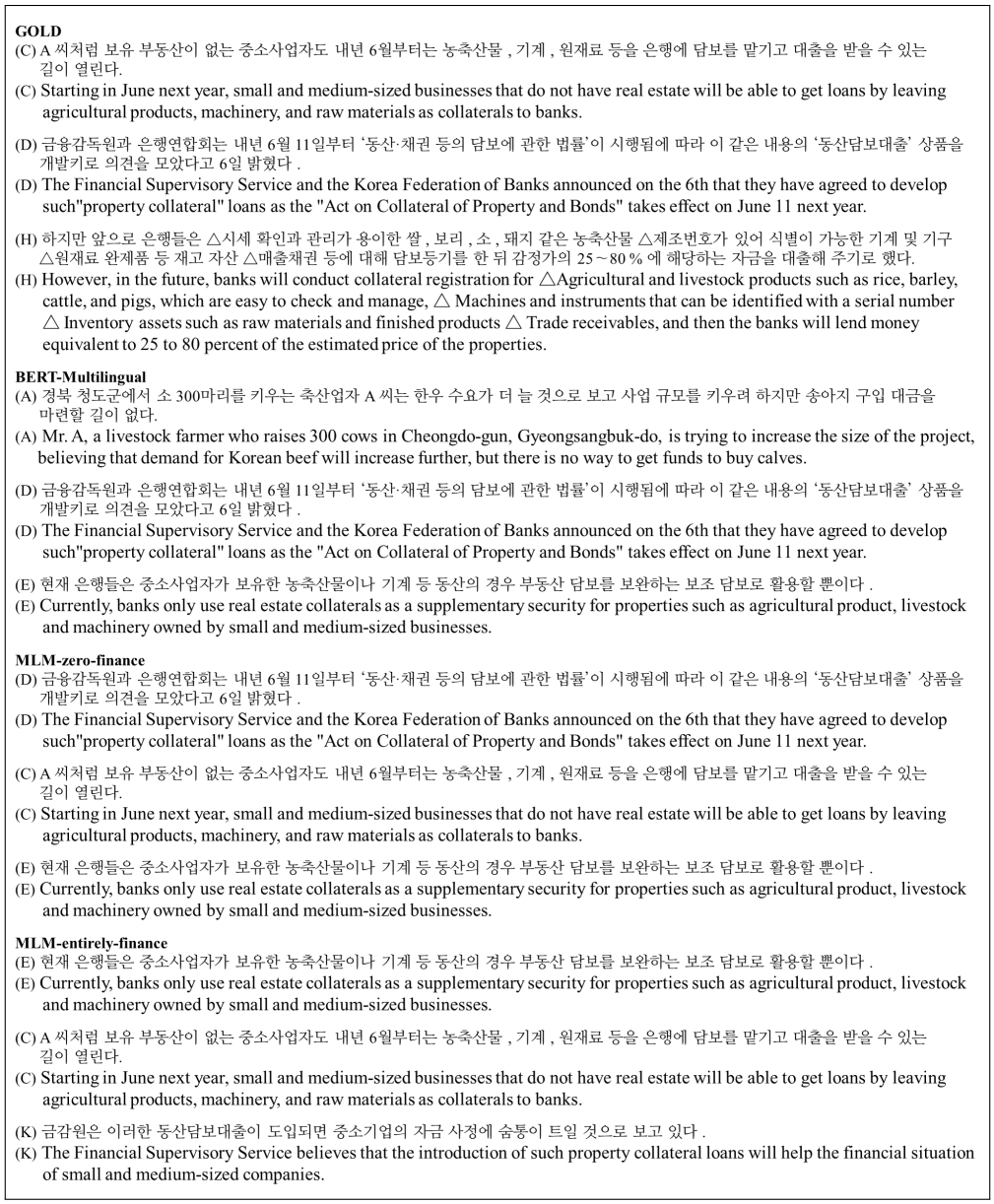


Figure 5: Summarization results of the models on a sample article. Each sentence in the example is shown with an English gloss on the line below.

mance, followed by MLM-zero-finance. KoBERT and BERT-Multilingual showed competitive performance, and BERT-Multilingual showed higher performance than that of BERT-Multilingual-plus-finance. We consider this result non-trivial, which shows the case where the DAPT's performance gain reported by Gururangan et al. (2020) is not observed.

As a further analysis, we investigated predicted sentence positions in news articles. In Figure 3, human-generated gold summaries are evenly distributed, whereas the summaries of BERT-Multilingual are about 39% of sentences selected in the first three sentences. The summaries of MLM-zero-finance and MLM-entirely-finance are more evenly distributed than that of BERT-Multilingual. We speculate that MLM-zero-finance and MLM-entirely-finance are less dependent on positional information as well as more sensitive to complex expressions with numbers appearing in the middle and latter parts of news articles.

**자금난 中企 숨통 트일 듯.**
**Small and medium-sized businesses with insufficient funds will be able to breathe.**

(A) 경북 청도군에서 소 300마리를 키우는 축산업자 A 씨는 한우 수요가 더 늘 것으로 보고 사업 규모를 키우려 하지만 송아지 구입 대금을 마련할 길이 없다.
(A) Mr. A, a livestock farmer who raises 300 cows in Cheongdo-gun, Gyeongsangbuk-do, is trying to increase the size of the project, believing that demand for Korean beef will increase further, but there is no way to get funds to buy calves.

(B) 은행에서 부동산 담보로만 대출을 해주는 데다 낮은 신용등급 때문에 신용대출도 어렵기 때문이다.
(B) This is because banks only provide loans as real estate collaterals, and credit loans are also difficult due to low credit ratings.

(C) A 씨처럼 보유 부동산이 없는 중소사업자도 내년 6월부터는 농축산물, 기계, 원재료 등을 은행에 담보로 맡기고 대출을 받을 수 있는 길이 열린다.
(C) Starting in June next year, small and medium-sized businesses that do not have real estate will be able to get loans by leaving agricultural products, machinery, and raw materials as collaterals to banks.

(D) 금융감독원과 은행연합회는 내년 6월 11일부터 '동산·채권 등의 담보에 관한 법률'이 시행됨에 따라 이 같은 내용의 '동산담보대출' 상품을 개발키로 의견을 모았다고 6일 밝혔다.
(D) The Financial Supervisory Service and the Korea Federation of Banks announced on the 6th that they have agreed to develop such a "property collateral" loans as the "Act on Collateral of Property and Bonds" takes effect on June 11 next year.

(E) 현재 은행들은 중소사업자가 보유한 농축산물이나 기계 등 동산의 경우 부동산 담보를 보완하는 보조 담보로 활용할 뿐이다.
(E) Currently, banks only use real estate collaterals as a supplementary security for properties such as agricultural products, livestock and machinery owned by small and medium-sized businesses.

(F) 부동산이 없으면 담보대출이 사실상 어렵다는 얘기다.
(F) In other words, collateral loans have been virtually difficult without real estate.

(G) 이 때문에 6월 말 현재 은행권의 전체 동산담보대출 규모는 747억 원으로 전체 기업 대출금(567조5000억 원)의 0.01%에 그치고 있다.
(G) Therefore, as of the end of June, the total amount of property-collateral loans in the banking sector remains at KRW 74.7 billion, accounting for 0.01% of total corporate loans (567.5 trillion KRW).

(H) 하지만 앞으로 은행들은 △시세 확인과 관리가 용이한 쌀, 보리, 소, 돼지 같은 농축산물 △제조번호가 있어 식별이 가능한 기계 및 기구 △원재료 완제품 등 재고 자산 △매출채권 등에 대해 담보등기를 한 뒤 감정가의 25~80%에 해당하는 자금을 대출해 주기로 했다.
(H) However, in the future, banks will conduct collateral registration for △Agricultural and livestock products such as rice, barley, cattle, and pigs, which are easy to check and manage, △ Machines and instruments that can be identified with a serial number △ Inventory assets such as raw materials and finished products △ Trade receivables, and then the banks will lend money equivalent to 25 to 80 percent of the estimated price of the properties.

(I) 대출 기간은 담보의 성격과 자금 용도에 따라 달라진다.
(I) The term of the loan depends on the nature of the collateral and the purpose of the funds.

(J) 농축산물 담보로 운전자금을 대출받을 때는 1년 기한의 만기 일시 상환 조건이며 기계류를 담보로 시설자금을 대출받는 사업자는 5년 기한으로 원금 분할 또는 만기 일시 상환 조건 가운데 선택해 대출받을 수 있다.
(J) When receiving a loan for working capital as collateral for agricultural and livestock products, it is a one-year lump-sum repayment condition, and businesses borrowing facility funds with machinery as collateral can choose between the one-time and lump-sum repayment conditions of the principal installment for a five-year term.

(K) 금감원은 이러한 동산담보대출이 도입되면 중소기업의 자금 사정에 숨통이 트일 것으로 보고 있다.
(K) The Financial Supervisory Service believes that the introduction of such property collateral loans will help the financial situation of small and medium-sized companies.

(L) 공장을 빌려 쓰는 대부분의 중소기업들은 경기가 부진할 때면 자금을 융통할 방법이 없어 재고나 설비를 쌓아둔 채 도산하는 사례가 적지 않지만 앞으로는 은행에 동산을 담보로 맡기고 재도전할 수 있기 때문이다.
(L) Most small and medium-sized companies that borrow factories often go bankrupt with stock or equipment piled up due to lack of ways to liquidate the properties when the economy is sluggish, but in the future, they can leave their property as collateral to banks and try again.

Figure 6: An example of financial news article from our corpus. Each sentence in the example is shown with an English gloss on the line below.

#### 5.2.2 Impact of Dataset Domain on Fine-tuning

Figure 4 shows the performance difference according to the domain shift of the training dataset where the size of the training dataset stays the same. It should be noted that, for the *0% finance* (in the X-axis), the training dataset is 432 news articles with a category other than the finance. For the *50% finance*, the training dataset is the summation of the 216 non-financial news articles and 216 financial news articles. For the *100% finance*, the training dataset contains the 432 financial news articles. It should be noted that the performance was measured against the test data for the financial news articles. MLM-entirely-finance showed 0.61 ROUGE-1, 0.43 ROUGE-2, and 0.53 ROUGE-L for the *100% finance* case. We find that the scores are lower than those of MLM-entirely-finance in Table 4. We see that non-financial news articles can be used to improve performance, provided that such near-target domain dataset is available for fine-tuning. The *100% finance* case is for the case where 432 training samples are used, while 3,040 training samples were used for the case in Table 4.

Provided that the dataset size is controlled to be

the same, MLM-entirely-finance showed the highest ROUGE-1, ROUGE-2, and ROUGE-L when only the financial news articles were used for fine-tuning. We see that the performance was best when the domains of datasets used in the pre-training and fine-tuning phases were related to the target task domain. MLM-zero-finance showed the best performance when no financial news articles were used for fine-tuning. We observe that the performance improved when the domains of datasets in the two phases were synchronized, which is similar to the observation of BERT-Multilingual-plus-finance with domain-adaptive pre-training. We believe that, when training with multiple datasets of different domains, the order in which the datasets are used for training could also be an important factor in model performance. We leave further details for future work.

#### 5.2.3 Case Study

Figure 5 shows an example of the automatic summarization results for the article in Figure 6.

We see that the phrase *property collateral loan* can represent the topic of the article. In the gold standard, all the sentences contain tokens that are one of the three tokens within the phrase. The first sentence (C) includes the word *collaterals*. The second sentence (D) includes the phrase *"property collateral" loans* and the phrase *Collateral of Property*. The third sentence (H) contains the word *collateral*. Moreover, phrases like *agricultural products, machinery, and raw materials* appear in the article, and we speculate that the writer (of the gold standard) knew that *property* and *collateral* are concepts that include agricultural products, machinery, and raw materials. Three sentences are in the gold standard. The first sentence (C) includes the phrase *agricultural products, machinery, and raw materials* and the third sentence (H) includes the phrase *Machines and instruments that can be identified with a serial number Inventory assets such as raw materials and finished products Trade receivables*.

For the summarization generated by BERT-Multilingual, the first sentence (A) is considered as more of a start to an anecdote where the anecdote is only related to the topic of the article. Conceptually it is true that *cow* and *calves* are under the category of *property*. Considering the context, they are stated as the product that Mr. A plans to produce and sell, not as the collateral that Mr. A will register for a loan. They are not considered as expressions that reveal the model's knowledge of the meaning of *property* and *collateral*. The second sentence (D) includes the phrase *"property collateral" loans* and the phrase *Collateral of Property*, and they have a direct overlap with the (presumed) topic phrase *property collateral loan*. Likewise, the third sentence (E) included the phrase *collaterals* and *properties*. The third sentence (E) also included the phrase *agricultural, livestock and machinery* that can be seen as being under the category of *collaterals* and *properties*.

For the summarization results generated by MLM-zero-finance and MLM-entirely-finance, all three sentences in each of the summary result contained the words that are directly overlapped with the presumed topic phrase *property collateral loan*. For the summary generated by MLM-zero-finance, the second sentence (C) and the third sentence (E) contained a phrase that indicates objects under the category of *property* and *collateral*, such as *agricultural products* and *machinery*. For the summary generated by MLM-entirely-finance, the first sentence (C) and the third sentence (H) contained such phrases.

## 6 Conclusion

In this paper, we presented a transformer-based model for extractive summarization, focusing on financial news articles in Korean. We analyzed the interplay between the domain of the dataset used for pre-training and the domain used for fine-tuning, and found that the model performs better when the domain of the pre-training dataset matches the domain of the fine-tuning dataset, which can be different from that of the target. To the best of our knowledge, this is the first study to analyze the impact of the domains used in the models on extractive summarization of financial news articles in Korean. We believe that the results may also provide insights into training procedures and data preparation strategies for implementing models that give summaries in low-resource languages.

## Acknowledgments

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (No. 20180005820041001, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

## References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. 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)*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Changki Lee Jaewon Jeon, Hyunsun Hwang. 2019. Two-step korean document summarization using bert. *The Korean Institute of Information Scientists and Engineers*.

Diederik P. Kingma and Jimmy L. Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Jia J. Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020. How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

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:1907.11692*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *CoRR*, abs/1611.04230.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Shashi Narayan, Joshua Maynez, Jakub Adamek, Daniele Pighin, Blaz Bratanic, and Ryan McDonald. 2020. Stepwise extractive summarization and planning with structured transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language

models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Kaichun Yao, Libo Zhang, Tiejian Luo, and Yanjun Wu. 2018. Deep reinforcement learning for extractive document summarization. *Neurocomputing*, 284:52–62.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.