

# Fine-tuning and Domain Adaptive Pretraining in Korean Legal Domain

전종원

College of Liberal Studies, 2021-18031

cjw107@snu.ac.kr

## 1 Introduction

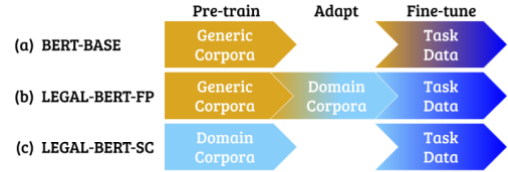
Pretrained language models based on Transformer have achieved great success in general use. Since these PLMs are pretrained on a large corpus, it is possible for the models to show great performance for downstream tasks through fine-tuning. However, there are still limitations when it comes to specialized domains because these models are pretrained using general corpus which does not include special vocabulary only used in these specific domains. Therefore, it is hard to expect these models to perform well simply by fine-tuning. This is also true for legal domains. Legal domains have their specialized vocabulary, and they have their own unique composition of logic.

In this paper, we explore several papers dealing with strategies in the legal domain and replicate the models using Korean Legal corpus and compare the performance. Although this paper does not propose either a totally new pretrained model or a fine-tuned model, note that this work can still contribute to understanding different state-of-the-art strategies for using pretrained language models in the legal domain.

## 2 Paper Summary

### 2.1 LEGAL-BERT: The Muppets straight out of Law School

The paper starts with the problem that BERT has been reported to under-perform in specialized domains, such as biomedical or scientific text. It points out that previous works do not investigate the effect of varying the number of pretraining steps, blindly adopt the hyper-parameter selection guidelines without investigation, and do not consider the effectiveness and efficiency of smaller models. To overcome these limitations, the paper explores three strategies: (a) use BERT out of the



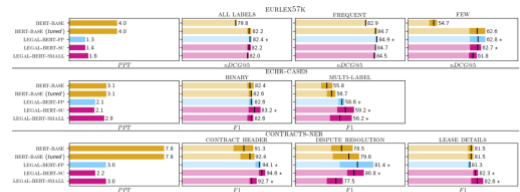
**Figure 1:** The three alternatives when employing BERT for NLP tasks in specialised domains: (a) use BERT out of the box, (b) further pre-train BERT (FP), and (c) pre-train BERT from scratch (SC). All strategies have a final fine-tuning step.

Figure 1. LEGAL-BERT strategies

box, (b) further pretrain BERT on domain-specific corpora, and (c) pretrain BERT from scratch on domain-specific corpora with a new vocabulary of sub-word units.

The key findings are: (i) Further pretraining (FP) or pretraining from scratch (SC) performs better than BERT-base. (ii) A broad hyper-parameter range can lead to substantially better performance. (iii) Smaller models can be competitive to larger ones in specialized domains. (iv) LEGAL-BERT released.

To be more specific, performance gains of FP and SC are stronger in the most challenging end-tasks such as multi-label classification. Also, LEGAL-BERT-SMALL is 3 times smaller but highly competitive to other versions of LEGAL-BERT.



**Figure 4:** Perplexities (PPT) and end-task results on test data across all datasets and all models considered. The reported results are averages over multiple runs also indicated by a vertical black line in each bar. The transparent and opaque parts of each bar show the minimum and maximum scores of the runs, respectively. A star indicates versions of LEGAL-BERT that perform better on average than the tuned BERT-BASE.

Figure 2. LEGAL-BERT results

When reading this paper and looking at the results, the biggest question I had was the similar performance between FP and SC. Since SC requires more time and resource to pretrain from scratch, I thought it would be quite obvious to expect a higher performance for LEGAL-BERT-SC. However, in tasks such as dispute resolution using CONTRACTS-NER dataset, FP model performed even better than SC. I believe this opens an area of further study to investigate the black box of pretrained language models and figure out the main components that affect the overall performance on specific datasets and tasks.

Also, if LEGAL-BERT-SMALL performs relatively well compared to its compact size, there might be area to allow smaller organizations to effectively develop their own models in an affordable price. If this can be commercialized, I believe this might open a new horizon to LLMs.

## 2.2 Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

Adaptive pretraining refers to further pretraining the pretrained model with the initialized weights. This paper performs domain-adaptive pretraining (DAPT) for four domains which are biomedical, computer science, news, and reviews. Also, it performs task-adaptive pretraining (TAPT) using task-specific data. In case where there is limited data, it performs TAPT after augmenting the task corpus.

Domain	Task	Additional Pretraining Phases			
		ROBERTa	DAPT	TAPT	DAPT + TAPT
BIO/MED	CHEMPROT	81.9 <sub>1.0</sub>	84.2 <sub>0.2</sub>	82.6 <sub>0.4</sub>	<b>84.4</b> <sub>0.4</sub>
	<sup>†</sup> RCT	87.2 <sub>0.1</sub>	87.6 <sub>0.1</sub>	87.7 <sub>0.1</sub>	<b>87.8</b> <sub>0.1</sub>
CS	ACL-ARC	63.0 <sub>6.8</sub>	75.4 <sub>2.5</sub>	67.4 <sub>1.8</sub>	<b>75.6</b> <sub>0.8</sub>
	SciERC	77.3 <sub>1.9</sub>	80.8 <sub>1.5</sub>	79.3 <sub>1.5</sub>	<b>81.3</b> <sub>0.8</sub>
NEWS	HYPERPARTISAN	86.6 <sub>0.9</sub>	88.2 <sub>5.9</sub>	<b>90.4</b> <sub>5.2</sub>	90.0 <sub>6.6</sub>
	<sup>†</sup> AGNEWS	93.9 <sub>0.2</sub>	93.9 <sub>0.2</sub>	94.5 <sub>0.1</sub>	<b>94.6</b> <sub>0.1</sub>
REVIEWS	<sup>†</sup> HELPLEFULNESS	65.1 <sub>3.4</sub>	66.5 <sub>1.4</sub>	68.5 <sub>1.9</sub>	<b>68.7</b> <sub>1.8</sub>
	<sup>†</sup> IMDB	95.0 <sub>0.2</sub>	95.4 <sub>0.1</sub>	95.5 <sub>0.1</sub>	<b>95.6</b> <sub>0.1</sub>

Table 5: Results on different phases of adaptive pretraining compared to the baseline ROBERTa (col. 1). On approaches are DAPT (col. 2, §3), TAPT (col. 3, §4), and a combination of both (col. 4). Reported results follow the same format as Table 3. State-of-the-art results we can compare to: CHEMPROT (84.6), RCT (92.9), ACL-ARC (71.0), SciERC (81.8), HYPERPARTISAN (94.8), AGNEWS (95.5), IMDB (96.2); references in §A.2.

Figure 3. DAPT and TAPT results

The key contribution of this paper is: (i) DAPT always improves performance, but adaptation to an irrelevant domain is likely to lower performance. (ii) Performing TAPT can improve performance for downstream tasks. (iii) Using both DAPT and TAPT shows better performance than only performing TAPT, but when there is enough labeled data, the difference is not big. (iv) Augmenting the task corpus using data selection

strategies to perform TAPT also improves performance.

## 2.3 KoLegal-BERT: Legal Language Representation Model for Legal Domain Text Mining

This paper simply proposes KoLegal-BERT, which implements the two strategies, pretraining from scratch and adaptive pretraining for legal domain. It uses open data released in AI Hub, news related to legal domain, and legal documents. The base model is klue-bert-base, with AdamW optimizer, and performs document classification task.

Model	Precision	Recall	f1-score
kobert-base	78.22	77.15	76.83
klue-bert-base	81.90	79.95	79.73
KoLegal-BERT-sc (small)	81.83	81.02	80.96
KoLegal-BERT-sc (base)	82.41	81.51	81.38
KoLegal-BERT-fp	82.89	82.14	82.08

표 4 언어모델에 따른 법률 문서 분류 모델 성능

Figure 4. Performance of KoLegal-BERT

KoLegal-BERT showed relatively higher performance than the base models. FP model showed the highest performance, and small SC model also performed well considering the relatively small size of data.

However, even though this paper argues that the improvement in performance indicates further possibility of using pretrained language models in legal domains, there remains a question mark whether the improvement can be seen as significant. The difference of performance is not that big, meaning that a higher-level strategy is needed for the implementation to be effective enough in the legal domain.

## 3 Experimental Setup

### 3.1 Overview

Based on the three papers above, it has been proven that pretraining pretrained language models can improve performance in specialized domains. However, there are some unsolved questions in previous works.

First, can we say that further pretraining and pretraining from scratch is effective enough compared to the resource needed to pretrain? Second, since Korean data in the legal domain is very limited, would legal tech companies find a way to survive in the industry using PLMs? In

other words, can simply obtaining more data and using the current strategies guarantee higher performance? In my opinion, since small size model pretrained from scratch showed relatively competitive performance, further studies are needed to figure out an effective way to improve performance in specialized domains with limited amount of data.

Although it is not easy for me to find answers for these problems due to lack of data and GPU resource, I've tried to replicate the existing strategies to find out if such methodology can still work in Korean legal data.

### 3.2 Task

Since pretraining from scratch is almost impossible for me to perform due to resource issues, I have narrowed down the project only to domain-adaptive further training. I tried to compare the results with the baseline model, and the fine-tuned one.

There are various tasks using the legal domain such as classifying case name based on case facts, classifying related statutes based on case facts, or summarizing precedents. I chose to work on classifying related statutes based on case facts considering the complexity of the task. According to various papers, it is known that domain adaptive pretraining is more effective for more difficult tasks. Since classifying case name seems too easy and summarization seems too difficult, I chose classifying statutes which seemed difficult enough to compare the results and find out the relative effectiveness of domain adaptive pretrained models. Also, I thought the difficulty of the project seemed sufficient for me to deal with.

Classifying statutes from case facts is a multi-label classification problem. Since there are a lot of existing statutes, and one case might have more than one related statute, it is different from, and more difficult than a simple binary classification task. One example of the statute classification problem is as the following.

Input:

"피고인은 2020. 6. 중순경 ... 절취하였다"

Output:

"형법 제 329 조"

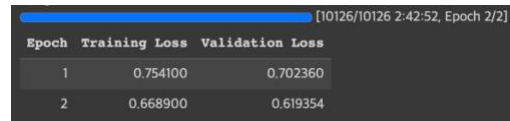
### 3.3 Dataset

I used the "LBox Open" dataset released by LBox, a legal tech startup in Korea. Since it is very hard

to find open dataset dealing with Korean legal texts, this was the one and only dataset I could easily use.

"LBox Open" dataset consists of four datasets which are precedent corpus, casename classification, statute classification, and summarization. I used the precedent corpus to perform domain adaptive pretraining and statute classification to perform finetuning and evaluation of the task.

The precedent\_corpus dataset has 150,000 precedents. Further pretraining the model using the whole dataset was impossible since the Colab blocked me from using too much GPU. Therefore, I used 30% of the entire corpus. Also, I was able to pretrain the model only for two epochs. Note that it took me three hours to complete the process. For convenience, I included the link in the ipynb file so that the saved pretrained model can be easily downloaded.



Epoch	Training Loss	Validation Loss
1	0.754100	0.702360
2	0.668900	0.619354

### 3.4 Models

I chose to use the KLUE-RoBERTa-base model because of the two following reasons.

First, I wanted to use a RoBERTa-base model rather than a BERT-base model. Since RoBERTa used larger data, pretrained longer, and trained with large batches, it is evaluated as a better trained general language model than BERT. Also, since previous works used BERT as their base models, using RoBERTa would give a point of difference to this project.

The second reason is because KLUE is the frequently used model which is trained with various data, and which has sufficient number of parameters. Since we are dealing with the legal domain, I thought it would be better to use a PLM trained with data from the news rather than other sources such as tweets or comments on the internet.

### 3.5 Code Structure

The code is implemented in three different sections, fine-tuning, domain adaptive pretraining, and the baseline model.

In all three models, KLUE-RoBERTa-base model and tokenizer is loaded using the AutoModel and AutoTokenizer function implemented in the Transformer library. Therefore, the dataset is tokenized using the auto-loaded tokenizer.

Also, since this is a multi-label classification, I implemented `id2label` and `label2id`, and a metrics function. Metrics function contains F1 score and `roc_auc`. F1 score is a harmonic mean of precision and recall, has a value between 0 and 1. The closer it is to 1, it represents better accuracy for the model. `Roc_auc` is the area under the roc curve. The higher the score, it means better performance of model.

For fine-tuning, I constructed data loader, used AdamW optimizer and defined training and evaluation function. For DAPT, I used the trainer function for convenience.

## 4 Experimental Results

By conducting statute classification task using the test datasets in the dataset, the performance result of the model for the given classification task is as below.

Model	Test	loss	F1	roc-auc
KLUE-RoBERTa-base	1	0.704	0.017	0.500
	2	0.360	0.018	0.500
Fine-tuned	1	0.039	0.071	0.577
	2	0.020	0.182	0.771
DAPT	1	0.037	0.157	0.732
	2	0.019	0.219	0.778

Figure 5. Performance Results

Fine-tuned model performed much better than the baseline model. Also, domain-adaptive pretrained model performed better than the fine-tuned model for both tests.

## 5 Conclusions and Future Work

We can conclude that as the previous papers suggest, in specialized domains such as the legal domain, further pretraining improves the performance. Fine-tuning might not be enough, for higher performance especially in complex tasks, further pretraining might be an effective way.

However, many limitations exist in this project.

First, this project simply replicates the existing papers. There is nothing special, except for the different baseline model and different dataset. One remarkable difference is that I used a RoBERTa based model with a Korean legal corpus, but this does not provide a significant new insight.

Second, due to lack of GPU resources, the result of this project cannot be seen as significant. Especially in the case of domain-adaptive

pretraining, out of 150,000 precedents, I could only use 30% of them, and pretraining happened only for two epochs. Also, in case of fine-tuning the batch size is 4 which is very small. At first, I tried with a size of 8, but the limited size of GPU stopped me from further implementations.

Third, it would have been better if I could have dealt with various tasks and compare the results between them. In this paper, I only considered multi-label classification task trying to classify the related statute given the case fact. If possible, by tackling other tasks such as precedent summarization, results can be much more enriched.

I believe the legal domain is an attractive field of study since the precedents and case facts are written by those who are capable of writing the most logical documents. Therefore, there are still room for further study to effectively apply language models in the legal domain to increase efficiency of the legal profession, and to allow the public to easily access legal domain. If more Korean legal dataset is constructed and better strategies are found, PLMs would lower the hurdle of the legal domain so that more people can easily have access to it, which is an important issue since law is something that is very much affined with our daily life.

## References

- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2898–2904, Online. Association for Computational Linguistics.
- 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 58<sup>th</sup> Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.
- Sangmin Park, Yejin Yoon, Jaeyun Lee, and Jaieun Kim. 2022. KoLegal-BERT: Legal Language Representation Model for Legal Domain Text Mining. In Proceedings of the 49<sup>th</sup> Korean Information Science Society Conference, pages 1061–1063, Online. Korean Institute of Information Scientists and Engineers.