Exploiting Language Characteristics for Legal Domain-Specific Language Model Pretraining

Inderjeet Nair and **Natwar Modani**
Adobe Research, India
{inair,nmodani}@adobe.com

Abstract

Pretraining large language models has resulted in tremendous performance improvement for many natural language processing (NLP) tasks. While for non-domain specific tasks, such models can be used directly, a common strategy to achieve better performance for specific domains involves pretraining these language models over domain specific data using objectives like Masked Language Modelling (MLM), Autoregressive Language Modelling, etc. While such pretraining addresses the change in vocabulary and style of language for the domain, it is otherwise a domain agnostic approach. In this work, we investigate the effect of incorporating pretraining objectives that explicitly tries to exploit the domain specific language characteristics in addition to such MLM based pretraining. Particularly, we examine two distinct characteristics associated with the legal domain and propose pretraining objectives modelling these characteristics. The proposed objectives target improvement of token-level feature representation, as well as aim to incorporate sentence level semantics. We demonstrate superiority in the performance of the models pretrained using our objectives against those trained using domain-agnostic objectives over several legal downstream tasks.

1 Introduction

Pre-trained language models exhibit superior performance in several NLP tasks. Most of the prominent language models optimized over Masked language modelling with BERT-like (Devlin et al., 2019; Liu et al., 2019b; He et al., 2020) architecture using large unlabelled corpus to achieve state of the art results across many NLP tasks. While the sentence-level tasks like paraphrase detection (El Desouki and Gomaa, 2019) and sentiment analysis (Zhang et al., 2018) benchmarks the capability of the model in effectively modeling the holistic representation of the sentence(s), the token-level tasks like named entity recognition (Li et al., 2020) attempted to assess the quality of contextualized token embeddings furnished by the models. However, direct application of these models to domain-specific downstream tasks yields sub-optimal performance (Lee et al., 2020), perhaps due to change in vocabulary and style of language.

To overcome this limitation, most commonly used approach involves pre-training a language model over domain specific corpora. For instance, PubMedBERT (Gu et al., 2021) and LEGAL-BERT (Chalkidis et al., 2020) achieved state-of-the-art results for the biomedical and legal domain specific tasks respectively by pre-training over domain specific corpus using a domain agnostic objective. In this paper, we argue that the performance of these models can be further improved by employing pre-training objectives that exploit the language characteristics of the domain. We examine two distinct language characteristics of the legal domain, propose pre-training objectives and finally demonstrate superior performance over domain-specific NLP tasks. Legal domain departs from the generic corpora in terms of specialized vocabulary, particularly formal syntax, domain-specific knowledge semantics, etc. to the extent that it can be classified as a distinct "sub-language" (Tiersma, 1999; Williams, 2007), which may be addressed by MLM based pretraining. In this paper, we study the following additional domain characteristics and formulate closely aligned objectives in addition to domain agnostic objectives like MLM:

1. **Templatized language:** Legal documents consist of clauses that are often derived from reusable text fragments with placeholders. The placeholders are substituted with appropriate replacements for specific documents. We include a pre-training objective for this characteristic by optimizing the model to distinguish the substitutions from the rest of the text. Since, there is no labelled dataset that provides such information, we also outline the

Findings of the Association for Computational Linguistics: EACL 2023, pages 2516–2526

process to approximately label the data-points with placeholder spans.

2. **Availability of Soft Labels:** Contracts and legally enforceable documents can be segmented into clauses which are sections defining terms and conditions and important provisions. Clauses can be categorized into distinct types based on the aspect they address, which (the categorization) may sometimes be available as a heading/title associated with the clauses. This categorization enables us to define semantic relations between clauses. For instance, clauses having same type are closer in meaning as compared to different typed clauses. This fact is instrumental in formulating an objective to obtain semantically aware holistic representation.

We leverage these two characteristics to design a pre-training strategy, and experimentally show that a language model pretrained using our strategy outperforms domain-specific language model which is trained only on domain-agnostic objectives, such as Masked Language Modelling.

The rest of the paper is organized as follows. In Section 2, we discuss some prominent frameworks that provisions domain specific pre-trained models and survey important works in the legal AI. In Section 3, we elucidate the details for the aforementioned legal domain characteristics and describe the objective formulation and dataset curation strategy. In Section 4 discusses the training. In Section 5, we briefly describe the baseline models used to compare the performance with our pretrained model for several legal domain tasks, and discuss the results. Finally, in Section 6, we conclude by explaining the implications of our work and discuss its natural extensions.

2 Related Works

2.1 Prominent domain adaptation pretraining approaches:

Pretrained language models trained over non-domain specific data such as transformers (Vaswani et al., 2017), BERT (Devlin et al., 2019) and its variants (Liu et al., 2019b; He et al., 2020) has resulted in state-of-the-arts results for several non-domain specific natural language processing downstream tasks. Owing to their success, a prominent approach to achieve superior results in domain-specific NLP tasks involves training these models over domain-specific corpora. For instance, to improve the performance of the models in biomedical downstream tasks, BIOBERT (Lee et al., 2020), Clinical BERT (Alsentzer et al., 2019), Clinical BIOBERT (Alsentzer et al., 2019) and PubMedBERT (Gu et al., 2021) were pretrained over speciality corpora closely associated with the biomedical domain using the MLM objective. Recently, (Chalkidis et al., 2020) proposed LEGAL-BERT, a language model pretrained using MLM over domain specific corpora, to achieve state-of-the-art performance for several legal downstream tasks. Most of these methods focus on choosing appropriate corpora for MLM pretraining and the selection of optimal hyperparameters in contrast to the approach taken in this work. Here, we propose a new direction to adapt a pretrained language model by utilizing language characteristics. In particular, by studying the language characteristics of the legal domain, we propose pretraining objectives that explicitly tries to learn these characteristics. While there are other approaches that adapts the language model to domain-specific tasks (Rietzler et al., 2020; Han and Eisenstein, 2019; Gururangan et al., 2020), our work mainly tries to address the problem of pretraining a language model for a particular domain.

2.2 Legal Artificial Intelligence (AI)

Legal AI refers to the application of AI/NLP techniques to solve several tasks in the legal domain (Zhong et al., 2020). Due to the distinct language characteristics of the legal domain, many legal domain-specific tasks requires the expertise of legal practitioners for solving them. Furthermore, the complexity of the associated tasks requires a significant time commitment even for experienced legal professionals. Thus, this motivated the development of legal AI to reduce the tedium in understanding and solving these legal tasks.

In the legal AI, task-specific methods and datasets were proposed for the following tasks: Legal Judgement Prediction (Aletras et al., 2016), Legal Entity Recognition and Classification (Cardellino et al., 2017), Legal Question Answering (Kim and Goebel, 2017), Automated Legal Review (Hendrycks et al., 2021), Legal Text Classification (Chalkidis et al., 2021), etc. Instead on improving task-specific solution approaches, our objective is to make improvements for several downstream tasks. The objective of this work in very

similar to that of (Chalkidis et al., 2020), however, our solution approach is very different.

3 Domain Specific Objectives

We now describe the legal domain characteristics which we will use for formulating the objectives. For each of the two objectives, we also describe the associated dataset used for training. We get different pretrained language model variants by incorporating various subsets of the following objectives while pretraining.

The process of coming up with the right set of domain specific language characteristics requires significant exposure to the domain. The authors have been investigating several legal domain natural language processing tasks, and have been interviewing several practitioners for an extended period of time. The insights are a result of reading many legal domain documents and the interactions with domain experts. For one to extend our ideas in other domains, we expect them to require similar long exposure to the domain in question and opportunities to interact with domain experts. While we believe that the two characteristics identified in this work are not unique only to the legal domain, one will need to carefully evaluate whether the same characteristics apply to their chosen domain as well.

3.1 Legal Domain as a Templating Language

Contracts include clauses which often use a standardized language with some placeholders which are substituted with appropriate values (e.g., names, dates, amounts, locations, etc) for specific contracts (Figure 1). These standardized fragments with placeholders are referred to as templates (Niemeyer and Knudsen, 2005) in software engineering parlance. We refer to the tokens in the template-generated clauses that remain common across contract documents as **static** tokens and the values filled into the placeholders as **dynamic** tokens.

We propose a pre-training objective that aims to detect the **dynamic** tokens/spans from text fragments in the legal documents. Using this objective, the language model can generate holistic representation for a text-fragment cognizant of the tokens forming the dynamic part and the tokens forming the static part. This can also result in better contextualized token representation for the task of named-entity recognition or other entity level tasks.

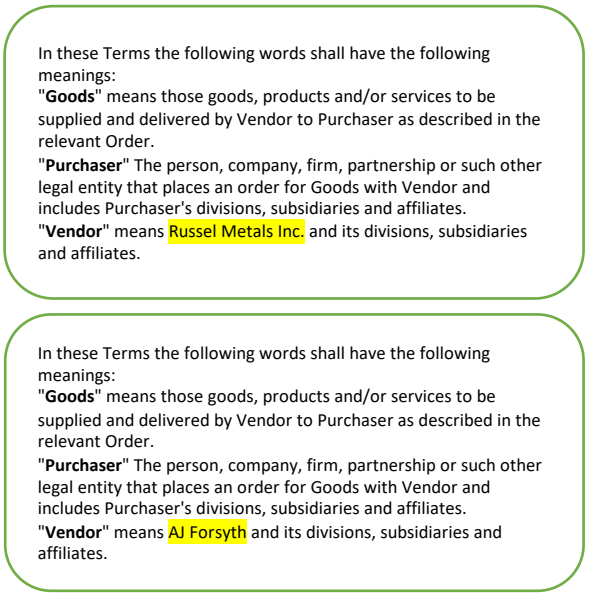


Figure 1: Clauses generated from same template: The above example is believed to be generated from a standardized clause template with a placeholder in place of the text in yellow highlight. Moreover the substituted text is observed to have close correspondence with organization named-entity.

3.1.1 Dataset

One of the challenges in utilizing this characteristic in the pre-training objective is the lack of any labelled dataset with such kind of information. To overcome this limitation, we propose a dataset curation strategy that provides data points with dynamic spans. The corpus to be labelled was formed by collecting all the clauses present in the LEDGAR dataset (Tuggener et al., 2020), which consists of over 700,000 provisions in contracts.

The data curation strategy mainly consists of two steps: a) **Grouping** clauses that have very high lexical similarity which are believed to be generated from a single underlying template, b) **Contrasting** data points in a pairwise fashion for every group to differentiate the dynamic part from the static using *google-diff-match-patch*[1]. Figure 3 illustrates the pipeline employed for annotating the dataset. Note that, while the contrasting tokens belong to *the dynamic part of the underlying text* (if the grouping was correct), there is inconclusive evidence for the rest of the tokens for considering them as *static* (For instance, Fig 2). This is due to the fact that some values can coincidentally be same for different instances of same clause, e.g.,

[1] https://github.com/google/diff-match-patch

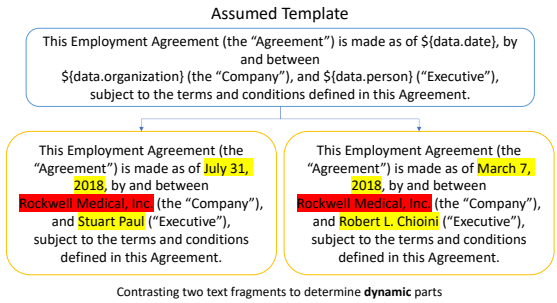


Figure 2: Limitation of the contrasting step: The two text fragments below belong to the same cluster and are believed to be generated from the template shown in the left. However, the process of contrasting annotates only some of the dynamic parts (highlighted in yellow) and misses out some (highlighted in red). Thus, the rest of the text should not be regarded as static in its entirety.

hiring date for two individuals can be the same, and therefore would not be marked as dynamic token by this strategy.

3.1.2 Objective Formulation

After applying the labelling strategy explained in the previous section, we obtain a token-wise labelled dataset $\mathcal{L} = \{(X_i, Y_i)\}_{i=1}^{M}$. A datapoint in $\mathcal{L}$ is a tuple (X_i, Y_i), where X_i represents a text-fragment as a sequence of tokens it contains ($X_i = [x_{ik}]_{k=1}^{|X_i|}$) and Y_i is the corresponding sequence of binary labels assigned to each token in X_i in the same order ($Y_i = [y_{ik}]_{k=1}^{|Y_i|}$ where $y_{ik} \in \{0, 1\}$), i.e. $y_{ik} = 1$ implies that x_{ik} belongs to the dynamic part and $y_{ik} = 0$ implies that the corresponding token can belong to any part.

Given such labelled dataset, we wish to train the language model $\mathcal{M}$ such that $\mathcal{M}_{\text{dyn}}(X_i, x_{ik})$ provides us with the likelihood of x_{ik} being dynamic. The subscript 'dyn' denotes the addition of task-specific overhead architecture for detecting **dynamic** spans. We cannot directly apply binary cross entropy objective over the token-level predictions as the negative labels in our case does not imply that the corresponding tokens *are* static. To overcome this obstacle, we use the framework of positive-unlabelled (PU) (Peng et al., 2019) learning where all the tokens associated with a positive label are regarded as dynamic and rest, associated with a negative label, are regarded as unlabelled. Under this framework, all the positively labelled tokens are collected with their parent text-fragment to form the set $\mathcal{X}_p = \{(X_i, x_i)\}_{i=1}^{n_p}$ where x_i represents a positively labelled token present in the text fragment X_i. This is also repeated for the negatively labelled datapoints to form the unlabelled set $\mathcal{X}_u = \{(X_i, x_i)\}_{i=1}^{n_u}$. PU learning optimizes the model parameters for the detection of dynamic parts by minimizing the following objective:

$$\mathcal{L}_{PU}(\mathcal{M}_{\text{dyn}}, \mathcal{X}_p, \mathcal{X}_u) = \frac{1}{n_u} \sum_{(X_u, x_u) \in \mathcal{X}_u} l(\mathcal{M}_{\text{dyn}}(X_u, x_u), 0) + \frac{\pi_p}{n_p} \sum_{(X_p, x_p) \in \mathcal{X}_p} (l(\mathcal{M}_{\text{dyn}}(X_p, x_p), 1) - l(\mathcal{M}_{\text{dyn}}(X_p, x_p), 0)) \quad (1)$$

where l is a positive-valued loss function that penalizes the distance between its arguments and $\pi_p \in [0, 1]$ is a hyperparameter. The above objective is derived from the following two terms: a term that incentivizes positively labeled instances to be classified as dynamic and a term that penalizes the unlabeled instances based on the assumption that the probability of being dynamic is equal to π_p Peng et al.. This implicitly assumes that the positive and unlabeled datapoints are sampled from the same distribution and the probability of a positive datapoint being labeled is independent of its input features. In contrast to the binary cross entropy objective, PU learning accounts for the possibility that some of the elements of $\mathcal{X}_u$ can be *dynamic*.

3.2 Soft Semantic Labels for Clauses

The legal essence of many contractual documents and agreements is formed by concatenating clauses which are crucial for defining terms and conditions and important provisions. These clauses can often be categorized which can be used to optimize the model to provide semantic-aware representation scheme, and sometimes, such categorization is available as a label/title with the clause text. Formally, we want to train the language model to learn a representation scheme that maps same category clauses from the data manifold onto metrically closer points in the mapped space. We believe that by infusing the ability to generate semantic-aware representation within model, the language model may offer better performance on sentence-level tasks.

3.2.1 Dataset

We used the LEDGAR Corpus (Tuggener et al., 2020) which is a collection of labelled legal clauses and provisions. This corpus was crawled from the contracts present in the website of U.S. Securities and Exchange Commission (SEC)[2]. While this dataset contains many clause instances with multiple labels, we retain only those clauses from this

[2] https://www.sec.gov/

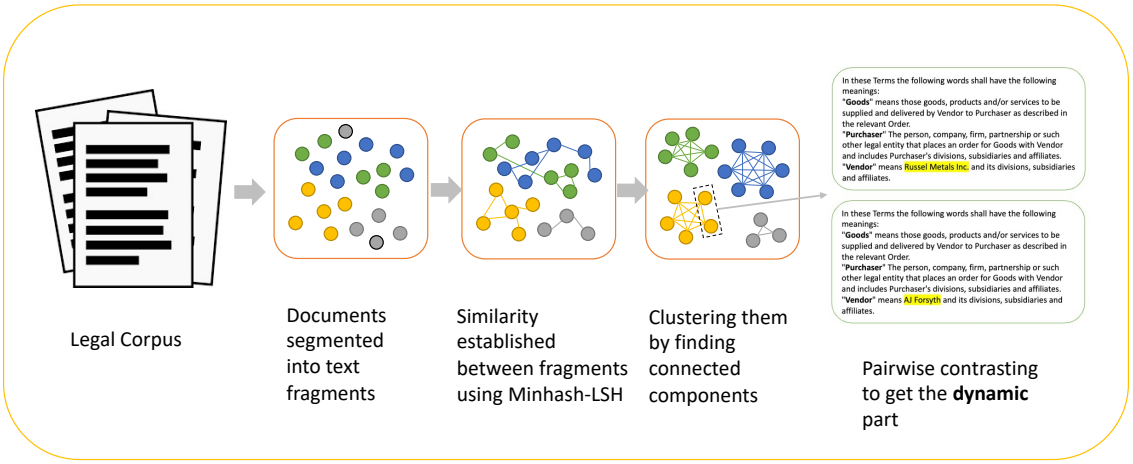


Figure 3: **Pipeline for dataset creation for dynamic part identification**: The clauses extracted from the LEDGAR corpus were originally obtained by segmenting legal documents into fragments. As clauses having fairly repetitive lexical structure are believed to be generated from the same template, the fragments are clustered using Minhash-LSH (Broder, 1997; Indyk et al., 1997), followed by finding the connected components. Finally, each pair in a cluster is contrasted to annotate what is dynamic among them.

corpus which are associated with a single label (roughly 83% of the dataset).

3.2.2 Objective Formulation

Given a language model $\mathcal{M}$, $\mathcal{M}_{\text{rep}}$ denotes task specific adaptation of the original language model to generate representation for a given sentence. We formulate our requirement as a task of metric learning where the goal is to learn a function $\mathcal{M}_{\text{rep}}(.) : \mathcal{X} \rightarrow \mathbb{R}^d$ that maps semantically closer input datapoints onto metrically closer points in $\mathbb{R}^d$. Here, $\mathcal{X}$ denotes the domain of input clauses / provisions. Under the triplet-loss formulation, every instance in the training dataset is a triplet (x_a, x_p, x_n) where the model tries to make the distance between the representations of x_a (*anchor*) and x_p (*positive*) smaller than that between x_a and x_n (*negative*) by atleast a margin m. Mathematically, the loss function l_{tri} is defined as follows:

$$l_{\text{tri}}(x_a, x_p, x_n) = [m + D(\mathcal{M}_{\text{rep}}(x_a), \mathcal{M}_{\text{rep}}(x_p)) - D(\mathcal{M}_{\text{rep}}(x_a), \mathcal{M}_{\text{rep}}(x_n))]_+ \quad (2)$$

In the above equation, $D(.,.) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ denotes a metric function measuring distances in the mapped space.

4 Training Details

We tune the parameters of our model using the algorithm employed for multi-task learning (Liu et al., 2019a). This framework optimizes the language model over multiple tasks. The language model is shared across different tasks by employing same encoder with shared parameters for all the task-specific overhead architectures. In each iteration of mini-batch gradient descent optimization, a task is randomly selected and corresponding task-specific mini-batch of data is sampled to apply single step of gradient descent using the task-specific objective. We curated the dataset for MLM pretraining by extracting text fragments from the SEC corpus as curated by Chalkidis et al., utilizing newline character ($\backslash$n) as the delimiter. In our ablation studies to understand the impact of various terms in the pretraining objective on downstream performance, we utilized a randomly selected subset of $40,000$ text fragments to quickly assess the importance of each of the terms. Thereafter, we also evaluate the performance of our model when a significantly larger corpora is provided for MLM.

In this paper, the parameters of the shared language model are initialized using the weights of LEGAL-BERT (12-layer, 768-hidden, 12-heads, 110M parameters)[3], a domain-specific language model pre-trained using MLM. Thereafter, we investigate the performance of the model variants listed in Table 1 by comparing against LEGAL-BERT. We do not assess the performance of non-domain specific models such as BERT (Devlin et al., 2019) as the superiority of LEGAL-BERT over BERT was demonstrated in (Chalkidis et al., 2020) for some of the legal downstream tasks.

[3]Distributed under CC BY-SA 4.0

Table 1: Model variants to be assessed in various legal downstream tasks (on top of LEGAL-BERT). Legal Corpus for MLM was collected by randomly sampling $40,000$ text fragments from the SEC corpus.

Model name	Description of Additional Pre-training
LB-PU	Dynamic span recognition using PU
LB-BC	Binary classification to identify dynamic tokens
LB-MLM	MLM over legal corpus
LB-PU-MLM	Multi-task training for PU and MLM over legal corpus
LB-TRI	Representation learning task using triplet margin loss
LB-TRI-MLM	Multi-task training for triplet margin loss and MLM over legal corpus
LB-PU-TRI	Multi-task training for PU and triplet margin loss
LB-PU-TRI-MLM	Multi-task training for PU, triplet margin loss and MLM over legal corpus

Table 2: Comparison between PU learning and Binary Classification for token-level tasks in terms of F_1-Scores (DPI: Dynamic Part Identification)

Model name	CUAD-NER	DPI
LEGAL-BERT	0.7040	0.7107
LB-PU	**0.7355**	**0.7507**
LB-BC	0.7221	0.6835

We used a 8 GPU A10G instance for training the models. While it took 32 hours to pretrain the model with best hyperparameter settings when only 40,000 datapoints for MLM is used, the model instance pretrained over the total SEC corpus (Chalkidis et al., 2020) consumed 800 hours. HuggingFace Transformers (Wolf et al., 2020) was used for both pretraining and experimental analysis.

5 Results and Discussion

We begin this section by validating the choice of using PU learning for dynamic part detection instead of binary token classification objective. In the subsequent subsection, we describe various legal downstream tasks and their associated data to be used in comparing the performance of the models in Table 1. As our models are derived from LEGAL-BERT, it is used as a baseline in our empirical analysis and we demonstrate the improvement of our model over it for several downstream tasks.

5.1 PU learning Versus Binary classification

5.1.1 Impact on downstream performance

In this subsection, we compare the performance of the model additional pretrained using PU learning (LB-PU) and binary classification (LB-BC) for named entity recognition (NER) and dynamic part identification (DPI).

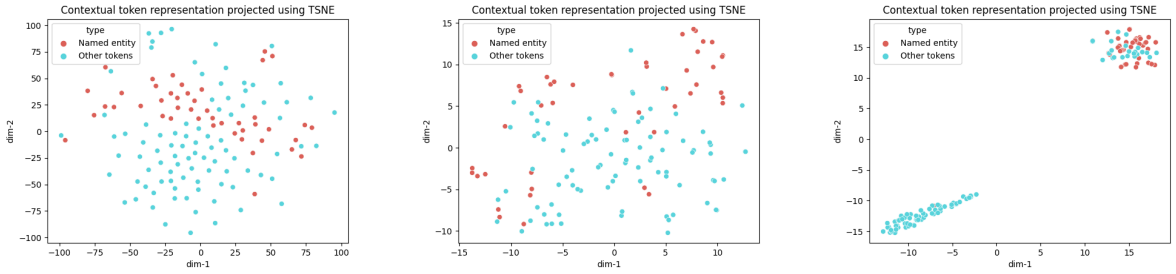
We use the NER adaptation of the Contract Understanding Atticus Dataset (CUAD) (Hendrycks et al., 2021). CUAD labels the *contracting-party* associated with each contract. This is used for constructing a NER dataset with *contracting-party* span annotations for each datapoint. This dataset consists of 16,636 training, 2,000 validation and a 10,000 testing samples.

As the dataset curated for pretraining the language model for dynamic part identification was approximately labeled, we manually annotated few text fragments by specifying the dynamic spans using the definition in section 3.1. This manual annotation furnished 132 training instances, 32 development instances and 50 testing instances. The performance was reported by computing the F_1-Score between the inferred spans and the ground truth dynamic spans.

The results shown in Table 2 justifies the utilization of PU learning objective. Our hypothesis that training the model to identify dynamic spans will improve its ability in recognizing named entities has been validated by the improvement in NER performance achieved through the use of the PU learning objective. This is further validated in the subsequent section through an examination of the feature representations generated by the model trained with/without PU learning. For the subsequent analysis, we disregard any models trained using binary classification objective owing to the results shown in the table. The decrease in the performance from LEGAL-BERT to LB-BC for NER and DPI stems from the fact that a subset of negatively labelled tokens in some instances are labelled as dynamic for other instances. This confuses the model in learning correct characteristics associated with these tokens, resulting in poor token-level representation.

Table 3: Performance for various legal domain task given in terms of F_1-Scores for CUAD-NER and DPI tasks, mean of F_1-Scores for MULTI-EURLEX tasks for Level 1, 2 and 3, and soft F_1-Score for Contract-Discovery task (Averaged for 5 runs).

Model name	CUAD-NER	DPI	MULTI-EURLEX	Contract-Discovery
LEGAL-BERT	0.7040	0.7107	0.7535	0.4591
LB-MLM	0.7344	0.7098	0.7525	0.4367
LB-PU	0.7355	0.7507	0.7488	0.0394
LB-PU-MLM	0.7427	0.7509	0.7451	0.1701
LB-TRI	0.7325	0.7380	0.7566	0.4979
LB-TRI-MLM	0.7462	0.7091	0.7567	0.5051
LB-PU-TRI	0.7320	0.7454	0.7513	0.5032
LB-PU-TRI-MLM	**0.7479**	**0.7628**	**0.7574**	**0.5119**



(a) LEGAL-BERT (NMI: 0.1837) (b) LB-BC (NMI: 0.0021) (c) LB-PU (NMI: 0.3812)

Figure 4: t-SNE projections of the contextualized embeddings obtained from different representation schemes. LB-PU visually performs the best in terms of segregating the named entities from the rest of the tokens.

5.1.2 Better feature representation for extracting named entities

We provide a qualitative justification for PU learning leading to better representation for extracting named entities in this subsection. In this assessment, we extract 30 text sentences from the **CUAD-NER** dataset that contain at least one named entity within it and compute the contextualized embeddings for the tokens in it using LEGAL-BERT, LB-BC and LB-PU. Thereafter, these embeddings are mapped to two dimensional manifold using t-SNE (Van der Maaten and Hinton, 2008) algorithm. Note that, we compute the embeddings using different representation schemes without fine-tuning on **CUAD-NER** to understand the impact of our token-level objective for distinguishing named entities from the rest of the tokens.

From Figure 4, we observe that the embeddings of the named entities and other tokens are not very well separated for LEGAL-BERT and LB-BC. On the other hand, LB-PU leads to much better segregation despite not being explicitly trained for the task of named entity recognition. This can be attributed to the observation that the dynamic part of a legal text fragment corresponds to a named entity most of the times. Since, LB-PU is explicitly pretrained for the task of dynamic part detection, it furnishes suitable representation scheme for segregating named entities. While LB-BC is trained for this task, it yields suboptimal representation scheme as it does not consider the possibility that some of the unlabelled tokens may be dynamic.

5.2 Results in various downstream tasks

Apart from **CUAD-NER** and **DPI** introduced in sec 5.1.1, we consider following additional tasks to compare the performance of different models:

1. **MULTI-EURLEX** (Chalkidis et al., 2021): This dataset is meant to assess the performance in the task of Large-Scale Multi-Label Text Classification (LSMTC). The datapoints in this dataset are curated from European leg-

islative documents (EUR-LEX) and the labels a derived from EUROVOC, a set of 4.3K European vocabulary labels. This dataset includes a total of 65K datapoints with the train-test-validation split of 55K-5K-5K respectively and involves fine-grained categorization of the label-set into 8 levels based on their hierarchy. We compute the performance of the model variants for 'level 1' (21 labels), 'level 2' (127 labels) and 'level 3' (567) (but report only the mean of these due to space constraint) as the other levels are not publicly available.

2. **Contract-Discovery** (Borchmann et al., 2020): This dataset is used to measure the performance of a model in semantic retrieval, where the task is to retrieve a span from a target document given a few examples (1 to 5) of similar clauses. The dataset uses about 600 target documents and is divided into 2 splits: development and test. Each of these splits consists of 5000 datapoints. The performance is evaluated by computing soft F_1 metric (Graliński et al., 2019) on the character-level inferred spans, which rewards proportionally to the extent of overlap between predicted and ground truth character spans. To solve this problem, we use the unsupervised method proposed by the authors of this task (Borchmann et al., 2020).

We can see from the Table 3 that the model pretrained using domain specific objectives achieves better performance than LEGAL-BERT for all the tasks. The models pretrained using only PU (LB-PU and LB-PU-MLM) only improves the performance for token-level tasks like CUAD-NER and DPI and achieves poor performance for other tasks. As these models only involve objectives at the token level, they offer inferior representations at the level of sentences / text-fragments as compared to other models which explains the poor performance in tasks like MULTI-EURLEX and Contract-Discovery. A similar effect is also observed for LB-MLM, where the model exhibits superior performance for some of the token-level tasks but exhibits poor performance for sentence level objective when compared against LEGAL-BERT as it does not involve any objective at the level of sentences. The models trained using triplet objective only (LB-TRI and LB-TRI-MLM) achieves better performance than LEGAL-BERT for all the tasks. This justifies the inclusion of the objective for learning semantic-aware representation scheme. We also observe that, inclusion of MLM for the model variants almost always improves the downstream performance. This indicates the usefulness of having domain-agnostic objective like MLM in the overall objective. The model pretrained using all the objectives (LB-PU-TRI-MLM) achieves best / competitive performance for most of the tasks. It is noteworthy that even though the objective of PU learning has no direct relation to tasks such as Contract-discovery and MULTI-EURLEX, the inclusion of PU learning in combination with Triplet loss and MLM leads to further improvement in the model's effectiveness in those tasks.

These results also emphasize the importance of MLM apart from the domain-specific objectives. Here, the pretraining over MLM was performed over a dataset with about $40,000$ text fragments. We believe that the performance of these models can be significantly improved by including a sufficiently larger dataset for MLM pretraining which is validated in the next subsection.

5.3 Performance when the size of MLM corpora is varied

In this section, we assess the performance of our model trained using the three objectives (PU + TRI + MLM) when the number of datapoints in the MLM corpus is varied. While the experiment performed in the previous subsection comprised of only 40,000 text fragments, this analysis assesses the model performance when the number of text fragments is varied from 1% to 100% of the total SEC corpus (Chalkidis et al., 2020).

The results shown in Table 4 clearly demonstrate that the downstream performance improves with the number of datapoints in the MLM corpus. Note that, **the pretraining corpus for LEGAL-BERT already comprises of the SEC corpus used in our analysis**. This fact also confirms the importance of involving the two objectives along with MLM for getting improved performance.

6 Conclusion

In this paper, we demonstrated a novel approach to enhance the performance of domain-specific language model across several specialty downstream tasks by exploiting the language characteristics. The objectives presented in this paper may not be applicable to all domains, which is a limitation

Table 4: Performance for various legal domain task given in terms of F_1-Scores for CUAD-NER and DPI tasks, mean of F_1-Scores for MULTI-EURLEX tasks for Level 1, 2 and 3, and soft F_1-Score for Contract-Discovery task when the number of datapoints in the MLM corpus is varied.

Number of training datapoints for MLM	Fraction of the overall SEC corpus	CUAD-NER	DPI	MULTI-EURLEX	Contract-Discovery
40,000	5.56×10^{-4}	0.7479	0.7628	0.7574	0.5119
720,000	0.01	0.7483	0.7651	0.7546	0.5210
7,200,000	0.10	0.7518	0.7662	0.7547	0.5145
18,000,000	0.25	0.7457	0.7636	0.7471	0.5158
72,000,000	1.00	**0.7523**	**0.7721**	**0.7577**	**0.5216**

of our work, but the idea of formulating objectives for learning domain-specific characteristics can be applied to other specialty domains (biomedical, programming languages, etc.). Future work might involve studying other characteristics of the legal domain and understanding their impact in downstream performance. We justified the positive impact of such pretraining across several downstream tasks by conducting extensive quantitative analysis.

We conclude this section by enumerating the natural extensions of this work for future:

1. In this work, we emphasized on two characteristics in the legal domain. However, the legal domain consists of several other domain-specific characteristics. For instance, the content in a legal agreement can be structured into different parts (preamble, recitals, list of clauses, etc) and the impact of involving a pretraining objective to infer the structure of a legal document on several tasks is yet to be understood. Thus, one line of future work may involve exhaustive study of language characteristics and understanding their influence in downstream tasks.

2. In the future, we plan to study the applicability of the introduced characteristics in other domains, such as programming languages where text fragments can be classified into categories like function blocks, variable declaration, etc. and contain both static and dynamic elements that can be templatized. This study may provide a thorough evaluation of the cross-domain applicability of these characteristics, including the assessment of their impact on downstream performance and the ease of curating relevant data. We would like to also motivate the researchers in applying the principle introduced in this paper for other domains (biomedical, finance, etc.). This necessitates careful investigation in order to extract domain-specific characteristics, as well as a mechanism for training the language model to understand these characteristics.

7 Limitations

We now discuss the limitations of our work. The first limitation (or requirement) is need for significant computational power. As we showed in Section 5.3 of our paper, when the corpus size for MLM training is increased from 0.0556% to 100% of the SEC corpus, while the performance improved by about 1% on multiple tasks, the computational requirement went up from 32 hours (on a 8 GPU A10G instance) to 800 hours.

Secondly, we had built our model on top of a domain specific pre-trained language model (which had used only MLM objective on a domain specific corpus). In theory, since we do include MLM as one of the objectives, we should be able to get comparable performance with or without domain specific pretrained language model. However, due to significant cost involved, we did not train a model starting from general domain language model (e.g., BERT or RoBERTa) to compare its performance against model built on top of domain specific pretrained language model. Therefore, we cannot make a claim if our proposed method would result in comparable performance improvement for the domains where such pre-trained models are not available.

Third, our method relies on identifying the domain specific characteristics and building objective functions suitable to exploit them. This requires building domain expertise and/or collaborat-

ing with domain experts. Since this process cannot be automated, it requires additional cost and human effort. Also, good automated data curation strategies may or may not be feasible for other domain specific characteristics, limiting using usefulness for training large language models.

Finally, we have only experimented with English language corpus. While the data curation strategy we used should be applicable in most other languages also for legal domain, the static/dynamic token classification task particularly may depend on grammatical rules for sentence construction, which may not be similar in all languages.

However, we believe that despite these limitations, our work points to possibility of improved performance of language models by using domain specific characteristics (beyond MLM based pretraining), which should open doors for more such explorations and significant advances in the state of art.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Łukasz Borchmann, Dawid Wisniewski, Andrzej Gretkowski, Izabela Kosmala, Dawid Jurkiewicz, Łukasz Szałkiewicz, Gabriela Pałka, Karol Kaczmarek, Agnieszka Kaliska, and Filip Graliński. 2020. Contract discovery: Dataset and a few-shot semantic retrieval challenge with competitive baselines. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4254–4268, Online. Association for Computational Linguistics.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 9–18.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohamed I El Desouki and Wael H Gomaa. 2019. Exploring the recent trends of paraphrase detection. *International Journal of Computer Applications*, 975(S 8887).

Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz Górecki. 2019. GEval: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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*, pages 8342–8360, Online. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *NeurIPS*.

Piotr Indyk, Rajeev Motwani, Prabhakar Raghavan, and Santosh Vempala. 1997. Locality-preserving hashing in multidimensional spaces. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 618–625.

Mi-Young Kim and Randy Goebel. 2017. Two-step cascaded textual entailment for legal bar exam question answering. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 283–290.

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):1234–1240.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Patrick Niemeyer and Jonathan Knudsen. 2005. *Learning java*. " O'Reilly Media, Inc.".

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuan-Jing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Peter M Tiersma. 1999. *Legal language*. University of Chicago Press.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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

Christopher Williams. 2007. *Tradition and change in legal English: Verbal constructions in prescriptive texts*, volume 20. Peter Lang.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.