DELTA: Pre-Train a Discriminative Encoder for Legal Case Retrieval via Structural Word Alignment

Haitao LI[1,2] Qingyao AI[1,2] Xinyan HAN[1,2], Jia CHEN[3], Qian DONG[1,2], Yiqun LIU[1,2*],

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2] Institute for Internet Judiciary, Tsinghua University, Beijing, China
[3] Xiaohongshu Inc
liht22@mails.tsinghua.edu.cn

Abstract

Recent research demonstrates the effectiveness of using pre-trained language models for legal case retrieval. Most of the existing works focus on improving the representation ability for the contextualized embedding of the $[CLS]$ token and calculate relevance using textual semantic similarity. However, in the legal domain, textual semantic similarity does not always imply that the cases are relevant enough. Instead, relevance in legal cases primarily depends on the similarity of key facts that impact the final judgment. Without proper treatments, the discriminative ability of learned representations could be limited since legal cases are lengthy and contain numerous non-key facts. To this end, we introduce DELTA, a discriminative model designed for legal case retrieval. The basic idea involves pinpointing key facts in legal cases and pulling the contextualized embedding of the $[CLS]$ token closer to the key facts while pushing away from the non-key facts, which can warm up the case embedding space in an unsupervised manner. To be specific, this study brings the word alignment mechanism to the contextual masked auto-encoder. First, we leverage shallow decoders to create information bottlenecks, aiming to enhance the representation ability. Second, we employ the deep decoder to enable "translation" between different structures, with the goal of pinpointing key facts to enhance discriminative ability. Comprehensive experiments conducted on publicly available legal benchmarks show that our approach can outperform existing state-of-the-art methods in legal case retrieval. It provides a new perspective on the in-depth understanding and processing of legal case documents.

Code — https://github.com/CSHaitao/DELTA

Introduction

Legal case retrieval, which focuses on retrieving relevant cases for a query case, is essential for supporting legal reasoning and decision-making (Shao et al. 2020; Li et al. 2023a,b). In recent years, neural retrieval models constructed with pre-trained language models (PLM), which are capable of capturing the latent semantics of text documents, have attracted significant attention in the field of legal case retrieval (Su et al. 2024). For instance, Xiao et al (Xiao et al. 2021) propose Lawformer based on Longformer. Li et al. (Li et al. 2023a) developed a structure-aware framework named SAILER which utilizes structural information in legal cases to better pretrain the text encoder.

Despite their success, existing pre-trained retrieval models for legal case retrieval are far from being perfect due to several problems. One particularly significant issue is their incapability of discriminating key legal facts, which are crucial for determining case relevance, from other texts that simply describe case background and unimportant facts. As highlighted in prior research (Ma et al. 2021; Shao et al. 2023), the relevance determination in legal case retrieval is complex and fundamentally different from general search relevance. Unlike open-domain retrieval which mostly relies on keyword matching or text semantic similarity, legal case relevance focuses more on the identification of key facts that are crucial for case decisions. Existing pre-trained models focus on constructing better representations that capture the text semantics of individual case documents. However, in the legal domain, better semantic representation vectors do not always lead to better discrimination of legal relevance if the representations focus on capturing facts that are unimportant from legal perspectives.

To address this issue, we propose a novel framework called DELTA, which stands for Pre-training a **D**iscriminative **E**ncoder for **L**egal Case Re**T**rieval via Structural Word **A**lignment. Inspired by SAILER, we further delve into the knowledge implied between different structures of legal case documents to pre-train the encoder. Specifically, DELTA employs an encoder-decoder architecture to achieve an in-depth understanding of legal cases and effective extraction of key information. Besides employing shallow decoders to create a bottleneck for the $[CLS]$ token and generate high-quality textual representations, DELTA also incorporates the Structural Word Alignment (SWA) task to identify key facts by "translating" the Fact section to the legal analysis and decisions in the Reasoning section. Furthermore, DELTA enhances the alignment between the case representation and its key factual information in the semantic space. This alignment is achieved by pulling the case representation closer to the key facts while simultaneously pushing it away from background information within the case document. The whole algorithm not only enhances the discriminative ability of the representation models from the

*Corresponding authors.

legal perspective but also helps legal users better track the key facts in retrieved cases, thus making the results more interpretable and trustworthy. To the best of our knowledge, DELTA is the first work to apply word alignment in legal case retrieval, which provides new perspectives for in-depth understanding and processing of legal cases. To validate the effectiveness of our model, we carried out comprehensive experiments on Chinese and English legal benchmarks. Empirical results indicate that DELTA significantly outperforms current state-of-the-art baselines.

Related Work

Legal Case Retrieval

Legal case retrieval has attracted considerable attention from both academia and industry. Researchers in legal case retrieval primarily focus on two categories of models: Expert Knowledge-based models (Zeng et al. 2005; Saravanan, Ravindran, and Raman 2009) and Natural Language Processing (NLP) models (Shao et al. 2020; Xiao et al. 2021; Chalkidis et al. 2020; Li et al. 2023a). Expert knowledge-based modeling enhances case representation by extending legal elements (Zeng et al. 2005) or developing ontological frameworks (Saravanan, Ravindran, and Raman 2009), including the introduction of new sub-elements to the legal problem element for a more comprehensive representation of legal cases. On the other hand, Natural Language Processing models employ deep learning techniques such as BERT and its variants to capture semantic similarities between cases at the text level. These models show great potential for legal case retrieval.

Recently, numerous studies have addressed the challenge of long texts in the legal field by segmenting the text into paragraphs or enlarging the inputs for language models. For instance, Shao et al. (Shao et al. 2020) segment legal documents into multiple paragraphs and subsequently aggregate the scores. Tang et al. (Tang et al. 2023) converted the cases into a Text-Attributed Case Graph to better represent legal instruments. Furthermore, many researchers have attempted to enhance performance by pre-training models on extensive legal corpora. For instance, the development of LEGAL-BERT (Chalkidis et al. 2020) involved collecting a significant array of English legal texts from various domains for pre-training. Li et al. (Li et al. 2023a) employed the specific structure of legal documents to pre-train SAILER, achieving state-of-the-art results on the legal dataset. Despite great success, the primary objective of these models is to enhance representing capabilities. However, in the legal domain, semantic similarity does not always correspond to case relevance. In this paper, our objective is to identify key facts within legal documents, thereby enhancing the discriminative ability of representation vectors.

Word Alignment in Language Translation

Word Alignment (WA) is a crucial component of statistical machine translation, primarily dedicated on establishing correspondences between words in a sentence pair (Li et al. 2019). This task plays a critical role in comprehending the relationships between languages and facilitating cross-language translation. In traditional statistical machine translation (Dyer, Chahuneau, and Smith 2013; Och and Ney 2003), word alignment usually requires an annotated parallel corpus, i.e., correspondences need to be manually created for each word in a sentence. This procedure is both labor-intensive and costly.

With the rapid evolution of Neural Machine Translation, alignment by attention brings a more flexible and efficient solution (Bahdanau, Cho, and Bengio 2014; Chatterjee et al. 2017; Li et al. 2019; Garg et al. 2019; Zenkel, Wuebker, and DeNero 2019; Liu et al. 2016). Neural machine translation can automatically learn word alignment from a corpus without manual annotation. Bahdanau et al. (Bahdanau, Cho, and Bengio 2014) were the first to demonstrate an example of word alignment using attention mechanisms in the RNNSearch model. Subsequently, Liu et al. (Liu et al. 2016) improved attention with the annotation results obtained by the statistical alignment tool, expecting better alignment results. Inspired by this work, we attempt to utilize word alignment between different structures to identify key facts in legal cases.

Task Description

The main purpose of legal case retrieval is to identify the relevant cases from the candidate cases for each query case. Formally, for a query q, the legal practitioner needs to find the top-k relevant cases from the candidate set $\mathbf{C}$. In the legal domain, these relevant cases $\mathbf{C}_q = c_1^*, c_2^*, ..., c_k^*$ are known as precedents, referring to historical cases that provide support for the judgment of the query case. In most legal case retrieval scenarios, q only contains the Facts component, while each candidate case containing the complete structure.

Generally speaking, a legal case usually consists of three parts: Fact, Reasoning, and Decision. The Fact section focuses on the argument, evidence, and basic facts. Since arguments and evidence are not all useful, Fact section usually contains a great deal of non-key facts. The Reasoning section reveals how the court selects and applies the legal rules. The Decision section is the definitive response of the court to the legal dispute. In practice, when legal practitioners draft a legal case document, they first analyze the key facts and subsequently construct the Reasoning section based on their experience. This process can be regarded as "translating" the Fact section into the Reasoning section. Numerous words or phrases within these two sections exhibit correspondences. Understanding the structures in legal cases and learning the correspondence between structures is essential to improve the performance of legal case retrieval.

Method

The framework of DELTA is shown as Figure 1. DELTA consists of three components, i.e., fact encoder, shallow decoders for different case sections, and a deep decoder for word alignment. The optimization goal is to make this dense embedding more expressive and distinguishable.

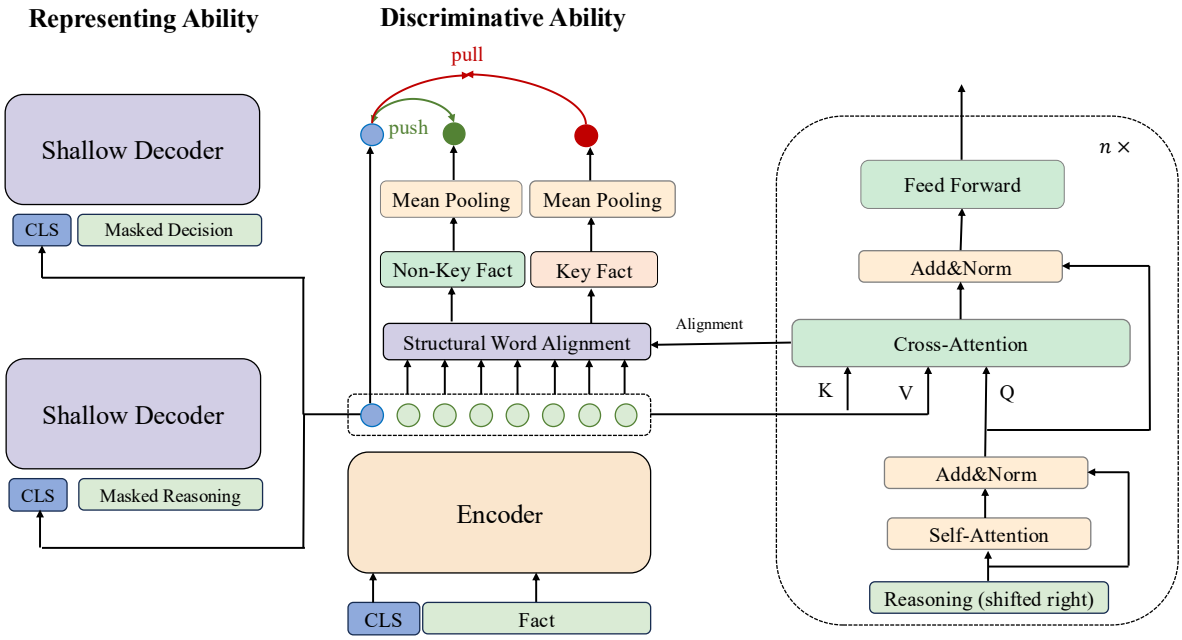

Figure 1: Pre-training designs of DELTA. DELTA creates information bottlenecks with two shallow decoders to improve representing ability of $[CLS]$ vector. Furthermore, Structural Word Alignment task is employed to identify key facts. DELTA pulls $[CLS]$ vectors closer to key facts and pushes them away from the non-key facts to enhance discriminative ability.

Encoder and Shallow Decoder

The Encoder encodes the Fact section into a high-quality representation vector to perform effective retrieval. In particular, the Fact section $F = [f_1, ..., f_n]$ is sampled, and then a portion of its tokens is randomly selected to be replaced with the $[MASK]$ token. To preserve enough information, only a small portion $(0 - 15\%)$ of the token is replaced. We define the masked tokens as $m(F)$ and input the masked fact F_{mask} to the encoder Ψ_F. We can get the contextualized embeddings of $[CLS]$ token h_F and other ordinary tokens H_F:

$$h_F, H_F = \Psi_F(F_{mask}) \tag{1}$$

The masked tokens are predicted using the standard approach of masked language modeling, resulting in the encoder's loss L_{mlm}.

As for decoding, we introduce two shallow decoders i.e., Reasoning Decoder and Decision Decoder. In this process, we concatenate the dense vector h_F with contextual sentence embedding and feed it into shallow decoders for text reconstruction. Specifically, we select some original tokens from the Reasoning section $R = [r_1, r_2, ..., r_n]$ and the Decision section $D = [d_1, d_2, ..., d_n]$ and replace them with $[MASK]$ token. Aggressive masking rates $(>= 30\%)$ are employed to provide adequate difficulty in reconstruction. Then, the original $[CLS]$ vectors in the Reasoning and Decision section are replaced with the dense representation h_F from Fact Encoder. The processed texts $m(R)$ and $m(D)$ are fed into their respective decoders. The loss for reconstructing the original text is defined as L_{dec}. As discussed in previous work (Lu et al. 2021), the shallow decoder has limited capabilities, so h_f is forced to represent more information for text reconstruction.

Structural Word Alignment

Many studies (Li et al. 2023a; Ma et al. 2023a) have already shown that the rich knowledge embedded in the structure of legal documents can effectively aid in model training. We further extend this spirit to improve the encoder's discriminative ability by exploiting potential correspondences between the different structures of legal case documents.

Since the Reasoning section discusses and analyzes all the key facts, each of which has a corresponding description, we attempt to identify the key fact tokens in the Fact section through the word alignment mechanism. However, annotated word alignment is challenging and labor-intensive. Instead, there has been some work exploring unsupervised word alignment with neural machine translation, which has proven to be effective (Garg et al. 2019; Zenkel, Wuebker, and DeNero 2019). Inspired by these studies, we propose the Structural Word Alignment (SWA) task in this section.

Specifically, given a Fact section $F = [f_1, ..., f_I]$ that serves as the source sentence and a Reasoning section $R = [r_1, r_2, ..., r_J]$ as the target sentence, where I and J represent the lengths of Fact and Reasoning section respectively, an alignment Λ is defined as a subset of the Cartesian product of the word positions (Och and Ney 2003):

$$\Lambda \subseteq \{(i, j) : i = 1, ..., I; j = 1, ..., J\}$$

The goal of the word alignment task is to establish a discrete alignment, representing a many-to-many mapping of source words to their corresponding translations in the target

sentence. To achieve unsupervised word alignment, DELTA introduces another decoder, with sufficiently deep layers to effectively perform the translation process. This decoder is autoregressive and its training objective can be formulated as:

$$L_{TLM} = -\sum_{j=1}^{J} \log p(r_j|F, r_{<j}, \theta) \tag{2}$$

where θ denotes the parameters of the deep decoder. Different from the shallow decoder, all ordinary tokens in the Fact section contribute to decoding the next token. It is also worth noting that this loss function does not optimize the parameters of the encoder to preserve its original encoding capability.

In this paper, we focus on guiding the cross-attention sublayer in the decoder to obtain the alignment relation. Formally, the representation of the i^{th} target token in the decoder as $q^i \in \mathbb{R}^{1*d_{emb}}$, where d_{emb} denotes the dimension of vectors. The output vectors of all source tokens from the encoder serve as the key matrix $K \in \mathbb{R}^{I*d_{emb}}$ and the value matrix $V \in \mathbb{R}^{I*d_{emb}}$. It is worth noting that the K and V matrices are identical, which corresponds to the outputs of the encoder. Then, the N heads project the query vector and the key and value matrices into distinct subspaces:

$$\tilde{q}_n^i = q^i W_n^Q, \tilde{K}_n = K W_n^K, \tilde{V}_n = V W_n^V \tag{3}$$

$$H_n^i = Attention(\tilde{q}_n^i, \tilde{K}_n, \tilde{V}_n) \tag{4}$$

$$M(q^i, K, V) = Concat(H_1^i, ..., H_N^i) W^O \tag{5}$$

where W_n^Q, W_n^K, W_n^V and W^O are all trainable parameters of the n^{th} head. $Concat(\cdot)$ denotes the concatenate of multi-head attention. The scaled dot-product attention is employed by each head:

$$Attention(\tilde{q}_n^i, \tilde{K}_n^i, \tilde{V}_n^i) = a_n^i \tilde{V}_n \tag{6}$$

$$a_n^i = softmax(\frac{\tilde{q}_n^i \tilde{K}_n^T}{\sqrt{d_{emb}}}) \tag{7}$$

where a_n^i represents the attention probabilities for the i^{th} target token across all source tokens in the n^{th} attention head. The multi-head attention mechanism of the transformer generates multiple attention matrices. To gain a deeper insight into the behavior of the encoder-decoder attention, we average the attention matrices across all heads within each layer to get a^i. The word alignment Λ can be readily extracted from the attention weight a^i according to the style of maximum a posterior strategy (MAP):

$$\Lambda_{i,j} = \begin{cases} 1 & j = \arg\max_{j'} a_{j'}^i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Following Garg et al. (Garg et al. 2019), we utilize the attentional probability of the penultimate layer, i.e., $l = L - 1$, as the alignment result, which is shown to provide the best alignment result in previous studies. Afterward, we group together the vectors a^i to obtain the attention matrix $A_{I \times J}$. The importance of tokens in the Fact section is calculated as follows:

$$X_i = \sum_{j=1}^{J} \Lambda_{i,j} A_{i,j} \tag{9}$$

According to X_i, we choose the top $p\%$ importance token in the Fact section as the key facts F_{key}, and the others are considered as non-key facts $F_{non-key}$. Then, we employ mean pooling on the embedding of these token to get their representations. In an ideal vector space for legal cases, $[CLS]$ embeddings are expected to be consistent with representations of key facts while far from non-key facts. Thus, contrastive learning is employed to achieve this objective, where representations of key facts are treated as positive examples h_p and those of non-key facts are negative examples h_n. The loss function L_{CON} is formulated as follows:

$$L_{CON} = -\log \frac{exp(sim(h_f, h_p))}{exp(sim(h_f, h_p)) + exp(sim(h_f, h_n))} \tag{10}$$

where $sim(\cdot)$ is the dot-product function. In practice, the in-batch negative technique (Karpukhin et al. 2020) is employed to better utilize positive and negative samples from the same batch. To be specific, given a case q, both positive and negative samples of the other cases in the same batch are considered as negative samples for case q. Notably, all positive and negative examples are obtained unsupervised without any manual annotation. This unsupervised text-level contrastive learning loss can effectively warm up the embedding space and improve the discriminative ability of the DELTA. Finally, the optimization objective of the model is the combination of the above losses, which is formulated as:

$$min. L_{mlm} + L_{dec} + \lambda_1 L_{TLM} + \lambda_2 L_{CON} \tag{11}$$

where λ_1 and λ_2 is the hyperparameters.

Compared to SAILER, DELTA cleverly leverages other ordinary tokens to more finely extract the legal knowledge embedded in different structures. The additional information provided by these ordinary tokens can help generate better semantic representations, offering greater potential for practical applications.

Experiment Setting

Datasets and Metrics

In this paper, we conducted experiments on both Chinese and English legal case retrieval benchmarks. Below, we provide detailed descriptions of these datasets:

- **LeCaRD** (Ma et al. 2021) is a widely used legal case retrieval dataset under the Chinese legal system. It contains 107 queries and 43,000 candidate case documents.
- **CAIL2022-LCR** is another Chinese legal case retrieval dataset, which has been provided as the test set for the CAIL2022 legal case retrieval competition. This dataset comprises 130 queries and 13,000 candidate cases.
- **COLIEE2022** (Kim et al. 2022) is an English dataset that serves as the official dataset for COLIEE2022 Task 1. This dataset consists of two parts: the training set and the test

set. The training set comprises 898 queries and 3,531 candidate cases, while the test set includes 300 queries and 1,263 candidate cases.

- **COLIEE2023** (Goebel et al. 2023) serves as the benchmark for evaluating legal case retrieval techniques in the COLIEE2023 competition. This dataset contains 959 query cases against 4,400 candidate cases for training and 319 query cases against 1,335 candidate cases for testing.

We follow the official metrics for these datasets to evaluate performance. Specifically, for LeCaRD and CAIL2022-LCR, we employ Precision@5, Recall@5, F1 score, NDCG@10 and NDCG@30. Furthermore, for the COLIEE datasets, we provide results for Precision@5, Recall@5, F1 score, R@100, and R@500.

Due to the limited number of queries and the lack of training sets in the Chinese datasets LeCaRD and CAIL2022-LCR, we only conducted zero-shot experiments on these datasets. For detailed statistics on the data, please refer to the Appendix.

Baselines

We conduct a comparison of DELTA with four distinct categories of baseline models, including Sparse Retrieval Models, General Pre-trained Models, Dense Retrieval Models, and Legal-oriented Pre-trained Models.

Regarding Sparse Retrieval Models, we consider taking BM25 (Robertson, Zaragoza et al. 2009) and QLD (Zhai 2008) as the baseline models, both of which are classical retrieval models based on lexical matching. Furthermore, General Pre-trained Models include BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019), with adaptations for the Chinese dataset using the corresponding Chinese versions of BERT and RoBERTa. We further extend our comparison to include a range of Dense Retrieval models: coCondneser (Gao and Callan 2021), SEED (Lu et al. 2021), COT-MAE (Wu et al. 2022). These models have retrieval-oriented optimization objectives and achieve state-of-the-art performance in web search tasks. To ensure equitable comparisons, we pre-train these models on the legal corpus with their optimal parameters.

Finally, we take into account pre-trained language models tailored to legal scenarios. These models comprise BERT_xs, Lawformer (Xiao et al. 2021), LEGAL-BERT (Chalkidis et al. 2020), SAILER (Li et al. 2023a), CaseEncoder (Ma et al. 2023b).

Since our work focuses on dense retrieval in legal scenarios, we did not include more time-consuming methods such as ColBERT (Khattab and Zaharia 2020), CaseGNN (Tang et al. 2023), CaseLink (Tang et al. 2024), and SLC (Ma et al. 2023a) in our baselines.

Implementation Details

Pre-training We initialize the encoder of the English/Chinese version of DELTA separately from bert-base-uncased/chinese-bert-wwm. The decoders are initialized from scratch. In line with Li et al. (Li et al. 2023a), the Chinese pre-training corpus comprises tens of millions of legal case documents, sourced from China Judgment Online [1]. We divide the case documents into three parts: Fact, Reasoning, Decision, with regular expression matching. Simple cases with facts less than 50 words are filtered. Similarly, the English pre-training corpus comprises an extensive collection of legal case documents from the U.S. federal and state courts [2]. The detailed data processing procedure is in the Appendix.

During the pre-training phase, we conduct training for up to 5 epochs using AdamW (Loshchilov and Hutter 2018) optimizer, with a learning rate of 5e-5, and a linear schedule with a warmup ratio of 0.1. The batch size is set to 72. The default mask ratio is set to 0.15 for the encoder and 0.45 for the shallow decoders. The shallow decoder is designed with a single transformer layer, while the deep decoder consists of six transformer layers. The hyperparameters p and is set to 60. A more elegant approach would be to adaptively assign different p values for each case, which we leave for future work. The hyperparameters λ_1 and λ_2 are both set to 1.

Fine-tuning During the fine-tuning phase, we discard all decoders and solely fine-tune the encoder. The training is performed on the COLIEE training set, employing a contrastive learning loss. For each given query, we employ BM25 to retrieve the top 100 related documents from the entire corpus, where irrelevant documents are treated as negative examples. The ratio of positive to negative examples is maintained at 1:15. We fine-tune the model up to 20 epochs employing the AdamW (Loshchilov and Hutter 2018) optimizer, with a learning rate set at 5e-6, a batch size of 4, and a linear schedule that includes a warmup ratio of 0.1. All experiments presented in this paper are conducted on 8 NVIDIA Tesla A100 GPUs.

Experiment Result

Zero-Shot Evaluation

We conduct zero-shot experiments on four legal benchmarks. The performance comparisons on two Chinese criminal law datasets are shown in Table 1. Due to space constraints, the zero-shot experiments on the English datasets are included in the Appendix. Based on the results in this table, we can derive the following conclusions:

- Without the guidance of training data, BM25 and QLD show competitive performance in legal case retrieval task, which is in line with previous findings (Li et al. 2023b). It shows that sparse retrieval methods are still very robust baselines in legal case retrieval task.
- General Pre-trained Models usually show suboptimal performance on legal case retrieval. This can be attributed to that their pre-training objectives are not tailored for similarity-based tasks, especially for legal case retrieval.
- Dense Retrieval model enhances their ability for similarity modeling, typically resulting in superior performance compared to General Pre-trained Models. However, even after being retrained on legal corpora, they still fall short

[1] https://wenshu.court.gov.cn/

[2] https://case.law/

Model	LeCaRD					CAIL2022-LCR				
	Precision	Recall	F1_score	NDCG@10	NDCG@30	Precision	Recall	F1_score	NDCG@10	NDCG@30
BM25	0.8916	0.1748	0.2922	0.7115**	0.8172*	0.8477**	0.2018**	0.3259**	0.7303**	0.8304*
QLD	0.8897	0.1737	0.2906	0.7157**	0.8373	0.8538**	0.2004**	0.3246**	0.7535**	0.8545
Chinese BERT	0.6654**	0.1263**	0.2123**	0.5252**	0.5374**	0.6600**	0.1487**	0.2427**	0.5604**	0.5618**
Chinese RoBERTa	0.8841	0.1778	0.2960	0.7438**	0.7897**	0.9046	0.2180	0.3513	0.8043**	0.8518
coCondenser	0.8411**	0.1648**	0.2756**	0.6719**	0.7404**	0.8138**	0.1931**	0.3121**	0.7063**	0.7607**
SEED	0.8411**	0.1575**	0.2653**	0.6721**	0.7330**	0.8369**	0.1906**	0.3105**	0.7365**	0.7815**
COT-MAE	0.8467**	0.1567**	0.2644**	0.6815**	0.7089**	0.8446**	0.1996**	0.3229**	0.7274**	0.7311**
BERT_xs	0.7159**	0.1377**	0.2309**	0.5695**	0.5751**	0.6462**	0.1369**	0.2259**	0.5236**	0.5206**
SAILER	0.9028	0.1902	0.3142	0.7979	0.8485	0.8938	0.2310	0.3671	0.8319	0.8660
Lawformer	0.8056**	0.1552**	0.2603**	0.6216**	0.6362**	0.7908**	0.1820**	0.2935**	0.6908**	0.6988**
CaseEncoder	0.8926	0.1879	0.3104	0.7850*	0.8391*	0.8947	0.2323	0.3688	0.8330	0.8670
DELTA	**0.9308**	**0.1959**	**0.3236**	**0.8117**	**0.8579**	**0.9077**	**0.2377**	**0.3767**	**0.8379**	**0.8709**

Table 1: Comparisons between DELTA and various baselines on Chinese benchmark on zero-shot setting. */** denotes that DELTA performs significantly better than baselines at $p < 0.05/0.01$ level using the Fisher randomization test (Rubin 1980). Best method in each column is marked bold.

Model	COLIEE2022					COLIEE2023				
	Precison	Recall	F1_score	R@100	R@200	Precison	Recall	F1_score	R@100	R@500
BM25	0.1307**	0.1552**	0.1418**	0.5866**	0.8416**	0.1222	0.2270	0.1589	0.6612**	0.8835**
QLD	0.1313**	0.1559**	0.1426**	0.6326**	0.8804**	0.1191*	0.2212*	0.1548*	0.7113*	0.9255*
BERT	0.1146**	0.1362**	0.1245**	0.5752**	0.8699**	0.0959**	0.1781**	0.1247**	0.6419**	0.9264*
RoBERTa	0.1524**	0.1805**	0.1650**	0.7517	0.9414	0.1003**	0.1862**	0.1303**	0.6967**	0.9070**
coCondenser	0.2393**	0.2842**	0.2598**	0.7508	0.9311	0.1197*	0.2223*	0.1556*	0.7355	0.9409
SEED	0.2266**	0.2692**	0.2461**	0.7527	0.9391	0.1223	0.2270	0.1589	0.7182*	0.9310*
COT-MAE	0.2427*	0.2882**	0.2634**	0.7608	0.9412	0.1229	0.2282*	0.1597	0.7347	0.9472
LEGALBERT	0.0713**	0.0847**	0.0774**	0.3432**	0.7123**	0.0545**	0.1013**	0.0709**	0.4659**	0.8629**
SAILER	0.2540*	0.3016*	0.2757*	0.7364*	0.9325	0.1253	0.2328	0.1629	0.7094**	0.9371*
DELTA	**0.2707**	**0.3214**	**0.2938**	**0.7636**	**0.9493**	**0.1316**	**0.2444**	**0.1711**	**0.7493**	**0.9502**

Table 2: Comparisons between DELTA and various baselines on English benchmark with fine-tuning setting. */** denotes that DELTA performs significantly better than baselines at $p < 0.05/0.01$ level using the Fisher randomization test (Rubin 1980). Best method in each column is marked bold.

Method	Precision	Recall	F1_score
DELTA	**0.0828**	**0.1526**	**0.1075**
w/o L_{CON}	0.0639	0.1187	0.0831
w/o L_{TLM}	0.0557	0.1036	0.0725
w/o L_{dec}	0.0608	0.1129	0.0790
w/o L_{mlm}	0.0796	0.1478	0.1035
w/o All	0.0338	0.0628	0.0440

Table 3: Ablation study on COLIEE2023 under zero-shot setting. Best results are marked bold.

in the legal domain due to a lack of understanding of legal cases.

- Legal-oriented Pre-trained Models are generally trained on extensive legal texts, which contributes to a better understanding of legal cases. However, the pre-training objectives of both BERT_xs and Lawformer do not focus on retrieval tasks, which limits their performance. SAILER, on the other hand, performs better than other pre-trained models, indicating that utilizing structural information in legal cases can lead to more effective case representations.
- Finally, We can observe that DELTA achieves the best performance on all metrics. Furthermore, it's worth noting that DELTA doesn't show a significant improvement compared to SAILER. This observation might be attributed to the nature of the Chinese legal system where cases with similar causes are typically considered relevant, rendering the need for discriminative ability less crucial. Nevertheless, by enhancing its discriminative capability, DELTA reaches the state-of-the-art in both Chinese datasets.

Fine-tuning Evaluation

We further compare DELTA with baselines on English benchmarks. For a fair comparison, we employ the same set of hyperparameters and fine-tuning data across various pre-trained models. As shown in Table 2, we have the following findings: (1) With the guidance of annotated data, the performance of pre-trained models is further improved. However, the majority of these models still face challenges in achieving satisfactory performance. (2) In comparison to COLIEE2023, pre-trained models a more substantial improvement on the COLIEE2022 dataset. This improvement could be attributed to the greater similarity between the training and testing sets in COLIEE2022. From this, We assume that robust unsupervised training approaches are crucial in legal scenarios where large-scale labeled data are lacking. (3) Overall, DELTA consistently achieves the best results on both datasets under fine-tuned evaluation. Our de-

signed unsupervised pretraining methods warm up the case-vector representation space, leading to better performance of DELTA with supervised data. This also indicates that the advantages of DELTA are universal, irrespective of the availability of annotated training data.

Ablation Studies

To better illustrate the effectiveness of our model design and pre-training tasks, we conduct ablation studies on COLIEE2023 in the zero-shot setting, which accurately represent the nature of the pre-trained model. The experimental results are presented in Table 3. From these results, we observe: (1) The removal of the L_{CON} component results in a significant degradation of performance, suggesting that the proposed contrastive learning loss plays a crucial role in learning discriminative representations. (2) The absence of L_{TLM}, which is crucial for training the deep decoder to provide accurate alignment, leads to a significant drop in performance. This highlights the importance of key facts in determining the relevance of legal cases. (3) Removal of L_{dec} also results in a decrease in model performance. This suggests that shallow decoders can create information bottlenecks, enhancing the representational capabilities of text vectors. (4) Consistent with previous research (Li et al. 2023a), the L_{mlm} task enhances the model's text comprehension, further boosting performance. These results demonstrate the effectiveness of our pre-training objectives. Both the representation and discriminative abilities of vectors are crucial for effective legal case retrieval.

Hyperparameter Analysis

In this section, we further investigate the impact of various components within the deep decoder. All reported results here are derived from experiments conducted on the COLIEE 2023 dataset in the zero-shot setting.

Impact of deep decoder layers number We initially study the influence of the number of deep decoder layers on performance. As shown in Table 4, performance consistently improves as the number of deep decoder layers increases, up to a point of 6 layers. Notably, a significant decline in performance is observed when the layer count is reduced to 2. We suspect that it is necessary for a certain depth in the decoder to more effectively execute the translation task. Overall, DELTA's performance in relation to its decoder layers exhibits notable robustness.

Impact of the ratio of key facts Subsequently, we explore the effect of the key fact ratio on performance. Specifically, this experiment involves grid searching the parameter p from 10% to 80%, in increments of 10%. As shown in Table 5. It is evident that the key fact ratio has a significant impact on model performance. At lower p values (e.g., 10%, 20%), the selected tokens fail to adequately represent the entire case, leading to reduced performance. Conversely, excessively high p values (e.g., 70%, 80%) may include non-key facts, potentially damaging performance. Overall, DELTA maintains commendable performance across a wide range of key fact ratios, with optimal performance achieved at a p value of 60%.

Decoder Layer	2	3	4	5	6	7
Precision	0.0645	0.0727	0.0771	0.0783	**0.0828**	0.0802
Recall	0.1199	0.1350	0.1431	0.1455	**0.1526**	0.1490
F1_score	0.0839	0.0945	0.1002	0.1018	**0.1075**	0.1043

Table 4: The impact of deep decoder layers number on COLIEE2023 under zero-shot setting. Best results are marked bold.

p	Precision	Recall	F1_score
10	0.0307	0.0570	0.0399
20	0.0501	0.0931	0.0651
30	0.0570	0.1059	0.0741
40	0.0658	0.1222	0.0855
50	0.0740	0.1373	0.0961
60	**0.0828**	**0.1526**	**0.1075**
70	0.0752	0.1397	0.0977
80	0.0708	0.1315	0.0920

Table 5: The impact of the ratio of key facts on COLIEE2023 under zero-shot setting. Best results are marked bold.

Visual Analysis

We employed t-SNE to visualize the vector distribution of legal case documents. This analysis was conducted on the COLIEE 2023 dataset in the zero-shot setting. Specifically, we visualized the sampled query and its top 200 candidate cases. The Case Study section in the Appendix presents the results for both SAILER and DELTA. For SAILER, the distribution of positive samples appears almost random, and the vector distribution is concentrated in a specific region. This is attributed to SAILER focus on modeling the representation of individual cases without considering the relationships between them. In contrast, DELTA addresses this limitation by utilizing the Structural Word Alignment task to warm up the vector space. As a result, in the vector space of DELTA, positive cases are more closely aligned with the query, and the overall vector distribution is more uniform. This approach allows DELTA's vector representation to more accurately reflect the relevance between legal cases. Overall, compared to SAILER, DELTA demonstrates a superior ability to generate discriminative case vectors.

Conclusion

In this paper, we present a novel pre-training framework DELTA for legal case retrieval. DELTA skillfully utilizes the translation process between different structures of legal case documents to identify key facts. These key facts are further employed to warm up the vector space, aiming to improve the discriminative ability of textual representations. Experimental results show that our pre-trained objectives contribute significantly to effective retrieval performance. Our method achieves state-of-the-art results on publicly available Chinese and English benchmarks. In the future, we will explore more ways to inject legal knowledge into pre-trained models for deeper understanding and analysis of legal case documents.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; and Androutsopoulos, I. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chatterjee, R.; Negri, M.; Turchi, M.; Federico, M.; Specia, L.; and Blain, F. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, 157–168.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dyer, C.; Chahuneau, V.; and Smith, N. A. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648.

Gao, L.; and Callan, J. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Garg, S.; Peitz, S.; Nallasamy, U.; and Paulik, M. 2019. Jointly learning to align and translate with transformer models. *arXiv preprint arXiv:1909.02074*.

Goebel, R.; Kano, Y.; Kim, M.-Y.; Rabelo, J.; Satoh, K.; and Yoshioka, M. 2023. Summary of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 472–480.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Khattab, O.; and Zaharia, M. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. arXiv:2004.12832.

Kim, M.-Y.; Rabelo, J.; Goebel, R.; Yoshioka, M.; Kano, Y.; and Satoh, K. 2022. COLIEE 2022 Summary: Methods for Legal Document Retrieval and Entailment. In *JSAI International Symposium on Artificial Intelligence*, 51–67. Springer.

Li, H.; Ai, Q.; Chen, J.; Dong, Q.; Wu, Y.; Liu, Y.; Chen, C.; and Tian, Q. 2023a. SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, 1035–1044. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394086.

Li, H.; Su, W.; Wang, C.; Wu, Y.; Ai, Q.; and Liu, Y. 2023b. THUIR@COLIEE 2023: Incorporating Structural Knowledge into Pre-trained Language Models for Legal Case Retrieval. arXiv:2305.06812.

Li, X.; Li, G.; Liu, L.; Meng, M.; and Shi, S. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1293–1303.

Liu, L.; Utiyama, M.; Finch, A.; and Sumita, E. 2016. Neural machine translation with supervised attention. *arXiv preprint arXiv:1609.04186*.

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

Loshchilov, I.; and Hutter, F. 2018. Fixing weight decay regularization in adam.

Lu, S.; He, D.; Xiong, C.; Ke, G.; Malik, W.; Dou, Z.; Bennett, P.; Liu, T.; and Overwijk, A. 2021. Less is More: Pretrain a Strong Text Encoder for Dense Retrieval Using a Weak Decoder. *arXiv preprint arXiv:2102.09206*.

Ma, Y.; Shao, Y.; Wu, Y.; Liu, Y.; Zhang, R.; Zhang, M.; and Ma, S. 2021. LeCaRD: a legal case retrieval dataset for Chinese law system. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2342–2348.

Ma, Y.; Wu, Y.; Ai, Q.; Liu, Y.; Shao, Y.; Zhang, M.; and Ma, S. 2023a. Incorporating Structural Information into Legal Case Retrieval. *ACM Trans. Inf. Syst.*, 42(2).

Ma, Y.; Wu, Y.; Su, W.; Ai, Q.; and Liu, Y. 2023b. CaseEncoder: A Knowledge-enhanced Pre-trained Model for Legal Case Encoding. arXiv:2305.05393.

Och, F. J.; and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1): 19–51.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Rubin, D. B. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American statistical association*, 75(371): 591–593.

Saravanan, M.; Ravindran, B.; and Raman, S. 2009. Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law*, 17(2): 101–124.

Shao, Y.; Li, H.; Wu, Y.; Liu, Y.; Ai, Q.; Mao, J.; Ma, Y.; and Ma, S. 2023. An Intent Taxonomy of Legal Case Retrieval. *ACM Trans. Inf. Syst.*, 42(2).

Shao, Y.; Mao, J.; Liu, Y.; Ma, W.; Satoh, K.; Zhang, M.; and Ma, S. 2020. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. In *IJCAI*, 3501–3507.

Su, W.; Ai, Q.; Wu, Y.; Ma, Y.; Li, H.; Liu, Y.; Wu, Z.; and Zhang, M. 2024. Caseformer: Pre-training for Legal Case Retrieval Based on Inter-Case Distinctions. arXiv:2311.00333.

Tang, Y.; Qiu, R.; Liu, Y.; Li, X.; and Huang, Z. 2023. CaseGNN: Graph Neural Networks for Legal Case Retrieval with Text-Attributed Graphs. arXiv:2312.11229.

Tang, Y.; Qiu, R.; Yin, H.; Li, X.; and Huang, Z. 2024. CaseLink: Inductive Graph Learning for Legal Case Retrieval. arXiv:2403.17780.

Wu, X.; Ma, G.; Lin, M.; Lin, Z.; Wang, Z.; and Hu, S. 2022. Contextual mask auto-encoder for dense passage retrieval. *arXiv preprint arXiv:2208.07670*.

Xiao, C.; Hu, X.; Liu, Z.; Tu, C.; and Sun, M. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2: 79–84.

Zeng, Y.; Wang, R.; Zeleznikow, J.; and Kemp, E. 2005. Knowledge representation for the intelligent legal case retrieval. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 339–345. Springer.

Zenkel, T.; Wuebker, J.; and DeNero, J. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*.

Zhai, C. 2008. Statistical language models for information retrieval. *Synthesis lectures on human language technologies*, 1(1): 1–141.