Fine-Grained Legal Argument-Pair Extraction via Coarse-Grained Pre-training

Chaojun Xiao[1,2†], **Yutao Sun**[1†], **Yuan Yao**[3], **Xu Han**[1], **Wenbin Zhang**[1]
Zhiyuan Liu[1,2*] **and Maosong Sun**[1*]
[1]NLP Group, DCST, IAI, BNRIST, Tsinghua University, Beijing
[2]Quan Cheng Laboratory
[3]National University of Singapore
xiaocj20@mails.tsinghua.edu.cn, {liuzy,sms}@tsinghua.edu.cn

Abstract

Legal Argument-Pair Extraction (LAE) is dedicated to the identification of interactive arguments targeting the same subject matter within legal complaints and corresponding defenses. This process serves as a foundation for automatically recognizing the focal points of disputes. Current methodologies predominantly conceptualize LAE as a supervised sentence-pair classification problem and usually necessitate extensive manual annotations, thereby constraining their scalability and general applicability. To this end, we present an innovative approach to LAE that focuses on fine-grained alignment of argument pairs, building upon coarse-grained complaint-defense pairs. This strategy stems from two key observations: 1) In general, every argument presented in a legal complaint is likely to be addressed by at least one corresponding argument in the defense. 2) It's rare for multiple complaint arguments to be addressed by a single defense argument; rather, each complaint argument usually corresponds to a unique defense argument. Motivated by these insights, we develop a specialized pre-training framework. Our model employs pre-training objectives designed to exploit the coarse-grained supervision signals. This enables expressive representations of legal arguments for LAE, even when working with a limited amount of labeled data. To verify the effectiveness of our model, we construct the largest LAE datasets from two representative causes, private lending, and contract dispute. The experimental results demonstrate that our model can effectively capture informative argument knowledge from unlabeled complaint-defense pairs and outperform the unsupervised and supervised baselines by 3.7 and 2.4 points on average respectively. Besides, our model can reach superior accuracy with only half manually annotated data. The datasets and code can be found in `https://github.com/thunlp/LAE`.

Keywords: Legal Argument-Pair Extraction, Coarse-Grained Pre-training

1. Introduction

Legal argument-pair extraction (LAE) aims to identify the interactive arguments with the same topic from the statements of the plaintiff and the defendant. During a trial process, the plaintiff and the defendant are supposed to state their *arguments* within the designated files called *complaint* and *defense*, respectively. An interactive *argument-pair* is defined as two interrelated arguments—one from the complaint and another from the defense—that address the same issue or topic. Arguments in interactive argument-pairs may either be in agreement, reinforcing similar points, or they could be in opposition, offering counterpoints to one another.

As illustrated in Figure 1, which features an example from a private lending case, both the plaintiff and the defendant find common ground on the reality of the lending transaction. However, they diverge significantly on other aspects, including the guarantor, the payment of interest, and the penalty. LAE seeks to facilitate the alignment of arguments presented in the complaint and defense documents. The task can aid legal professionals, especially judges, by highlighting the key factual disputes that require focused attention and providing a roadmap for subsequent investigation or analysis. In essence, LAE acts as a vital tool within legal assistant systems, aiming to improve the work efficiency of judges.

The task of argument-pair extraction has garnered increased interest recently, especially in the context of mining opinion interactions from dialogical argumentation. Most existing works formalize it as a sentence pair classification problem (Yuan et al., 2021a; Ji et al., 2021; Cheng et al., 2021; Bao et al., 2021b, 2022; Yuan et al., 2021b). However, these approaches rely heavily on substantial high-quality labeled data, which involves a time-consuming and labor-intensive manual annotation. Besides, these models often neglect the complex relationships within unlabeled complaint-defense pairs, resulting in suboptimal ability to LAE.

To address these issues, we introduce an argument-oriented pre-training framework for LAE. This framework leverages large-scale, coarse-grained complaint-defense pairs and is guided by two self-supervised supervision objectives, inspired by the following intuitive observations:

[†] Equal contribution.
[*] Corresponding authors.

LREC-COLING 2024, pages 7324–7335
20-25 May, 2024.

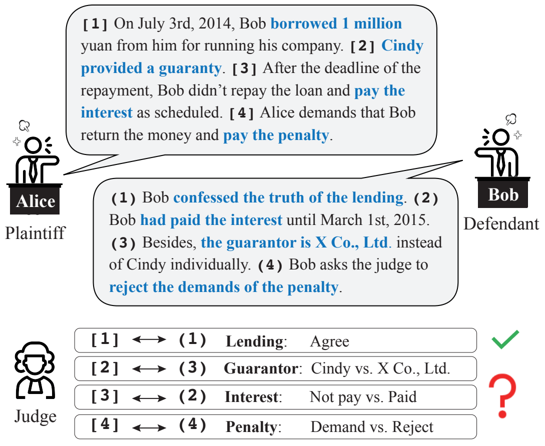
[1] On July 3rd, 2014, Bob borrowed 1 million yuan from him for running his company. [2] Cindy provided a guaranty. [3] After the deadline of the repayment, Bob didn't repay the loan and pay the interest as scheduled. [4] Alice demands that Bob return the money and pay the penalty.
Alice
Plaintiff
(1) Bob confessed the truth of the lending. (2) Bob had paid the interest until March 1st, 2015. (3) Besides, the guarantor is X Co., Ltd. instead of Cindy individually. (4) Bob asks the judge to reject the demands of the penalty.
Bob
Defendant
Judge
[1] ↔ (1) Lending: Agree ✓
[2] ↔ (3) Guarantor: Cindy vs. X Co., Ltd.
[3] ↔ (2) Interest: Not pay vs. Paid
[4] ↔ (4) Penalty: Demand vs. Reject
?

Figure 1: An example of the complaint and defense recorded in legal documents. The annotated argument-pairs are also presented, with the key phrases for each argument colored in blue.

1) Complaint-defense matching objective: Our research indicates that for many complaint arguments, there typically exists at least one corresponding defense argument that addresses them. Accordingly, we employ a contrastive learning strategy. For each complaint argument, the defense argument with the highest matching score within the corresponding defense is considered the positive instance. Meanwhile, arguments from unrelated cases serve as negative instances. This objective encourages each complaint argument to align closely with at least one defense argument. 2) Matching divergence objective: In a document, different complaint arguments are usually addressed by different defense arguments.Therefore, we aim to ensure that the matching distribution of defense arguments differs when comparing different complaint arguments. In other words, we encourage that different complaint arguments should not be associated with the same defense arguments. Our pre-training framework is independent from manually annotated data, and thus can be applied in large-scale self-supervised pre-training. After the pre-training phase, the model can serve in two capacities: as an unsupervised tool capable of directly extracting argument-pairs (unsupervised), or as a base model that can be further fine-tuned with a limited set of human-annotated data to improve its performance in a supervised manner (supervised).

To verify the effectiveness of our model, we construct two LAE datasets from the legal documents of two representative causes: private lending and contract dispute. Both two datasets contain tens of thousands of annotated cases and are the largest LAE datasets currently available. The evaluation results on these two datasets demonstrate that our proposed model can effectively leverage the underlying argument knowledge from large-scale legal documents, and achieve notable performance gains in both unsupervised and supervised settings. Remarkably, our model can achieve superior accuracy with only half labeled data, which further proves that our pre-training objectives can benefit models from the data scarcity problem.

2. Related Work

2.1. Argument Mining

Argument mining aims to analyze the semantic structures of argumentative text, such as debate dialog (Vecchi et al., 2021). It is a critical research field, which can benefit many downstream tasks, including opinion mining (Lawrence and Reed, 2019), stance detection (Küçük and Can, 2021). Argument mining has attracted growing attention, and existing research covers argument recognition (Trautmann et al., 2020; Grundler et al., 2022; Habernal et al., 2022), argument relation classification (Persing and Ng, 2016; Jo et al., 2021; Bao et al., 2021a), persuasiveness evaluation (Swanson et al., 2015; Khatib et al., 2020), argument summarization (Misra et al., 2015; Bar-Haim et al., 2020), etc.

Recently, some researchers begin to explore the argument-pair extraction task. Most existing methods address the task as a sentence pair classification problem and adopt various models for this task, such as discrete auto-encoder to capture information from varying aspects (Ji et al., 2021), multi-level attention layer to fuse passage-level information (Cheng et al., 2021), graph neural network to capture inter-sentence relationships (Bao et al., 2021b) or contrastive learning to identify valuable sentences (Shi et al., 2022). Some researchers adopt external knowledge, including the concept graph (Yuan et al., 2021b), and additional proposition detection task (Cheng et al., 2020), for further improvement. Notably, these works are parallel to ours, and due to the lack of external resources in the legal domain, we do not compare with these works in the experiments. Moreover, Yuan et al. (2021a) propose the first LAE dataset. These works ignore the underlying argument knowledge in unsupervised data, and require labor-intensive annotation.

2.2. Legal Artificial Intelligence

Legal artificial intelligence aims to empower legal tasks with artificial intelligence techniques and has received growing attention in recent years from natural language processing (NLP) researchers (Zhong et al., 2020; Bommasani et al., 2021; Xiao et al., 2023). With the booming development in NLP, many tasks have been proposed for automatic legal document analysis, such as legal judgment prediction (Zhong et al., 2018; Chalkidis et al., 2019), case retrieval (Shao et al., 2020; Ma et al., 2021; Yu et al., 2022), court view generation (Ye et al., 2018; Wu et al., 2020), information extraction (Chen et al., 2020; Yao et al., 2022). These tasks can provide handy references for legal practitioners and people seeking legal consulting.

Recent promising progress in pre-trained language models (PLMs) (Bommasani et al., 2021; Han et al., 2021) inspires legal AI scholars to dive into pre-training within the legal domain. Models trained on open-domain data exhibit suboptimal performance in legal tasks due to their limited understanding of legal specialized terminology. To address this issue, many researchers assemble extensive legal corpora to conduct continual pre-training based on open-domain PLMs (Chalkidis et al., 2020; Henderson et al., 2022; Cui et al., 2023; Xiao et al., 2021; Li et al., 2023). In this paper, we introduce a pioneering pre-training framework that seeks to leverage the argumentative structures inherent in legal documents.

Moreover, some researchers focus on legal argument mining, which aims to analyze the opinions of both parties in a case. Some researchers propose to detect the arguments in legal documents (Moens et al., 2007; Palau and Moens, 2009; Grundler et al., 2022; Habernal et al., 2022). Duan et al. (2019) summarize the lengthy and informative court debate via dispute focus mining. Different from previous work, we focus on the LAE and attempt to mine argument knowledge from the large-scale coarse-grained complaint-defense pairs, which can be applied easily to different scenarios.

3. Methodology

3.1. Notations

During the legal trial process, the plaintiff initiates the case by submitting a complaint that outlines their arguments. Subsequently, the defendant responds with a defense that either admits to or refutes the plaintiff's claims. LAE is designed to streamline this process by automatically matching related arguments that address the same topics from both the complaint and the defense, thus improving the work efficiency of judges in summarizing dispute focuses.

Although obtaining fine-grained annotations for individual argument-pairs is a labor-intensive task, acquiring coarse-grained complaint-defense pairs, denoted as (c, d), is considerably simpler. These pairs are often already well-structured and explicitly documented in legal records. In the context of a given complaint $c = q_1, ..., q_n$, each argument q_i (which we define as a sentence, following Ji et al. (2021)) aims to find its corresponding argument, denoted as p^+, from the defense $d = \{p_1, ..., p_m\}$ that discusses the same topic. Such an aligned pair (q_i, p^+) is then classified as an interactive argument-pair. For evaluation, we adopt the approach outlined in Yuan et al. (2021a). The model is presented with a set of 5 candidate defense arguments, consisting of 4 negative arguments (those that do not align with the complaint argument) and 1 positive argument (the one that aligns with the complaint argument). The model is then tasked with correctly identifying the positive candidate from this set.

3.2. Overall Framework

Our model is designed to harness the latent structure signals from unlabeled complaint-defense pairs to pre-train the model for LAE. Our model mainly relies on two intuitive observations: 1) Essentially, each argument presented in a complaint is expected to receive a corresponding response in the opposite defense. 2) Different arguments in complaints are likely to correspond to different arguments in defenses. Based on these observations, we propose two pre-training objectives: complaint-defense matching objective and matching divergence objective. It's worth noting that both objectives are symmetric and could be applied analogously to defense arguments as well. In the following sections, we will elaborate on these objectives primarily in the context of complaints for ease of explanation. The definitions for defense objectives can be obtained by a straightforward analogy. Figure 2 provides a graphical overview of the entire framework of our model. Our model takes a compliant argument as input and treats the corresponding defense as a positive bag of argument. The bag-level contrastive learning is performed by treating the argument with the highest matching score in the bag as the positive instance, while treating defense arguments from other cases as negative instances. Besides, we maximize the divergence between the matching distribution over defense arguments of different complaint arguments.

In the following, we first describe the model which maps argument pairs into matching scores. Then we introduce the two objectives inspired by the observations, which enforce the matching scores of related arguments to be high.

3.3. Argument-Pair Representation

In our architecture, we employ BERT (Devlin et al., 2019) as the foundational encoder for extracting argument-pair representations, as depicted in Figure 2. When we are presented with an argument pair (q, p), consisting of one argument from the complaint and another from the defense, we concatenate them with special tokens. The sequence $\{\texttt{[CLS]}, q, \texttt{[SEP]}, p, \texttt{[SEP]}\}$ is then input into the BERT encoder. The BERT encoder processes

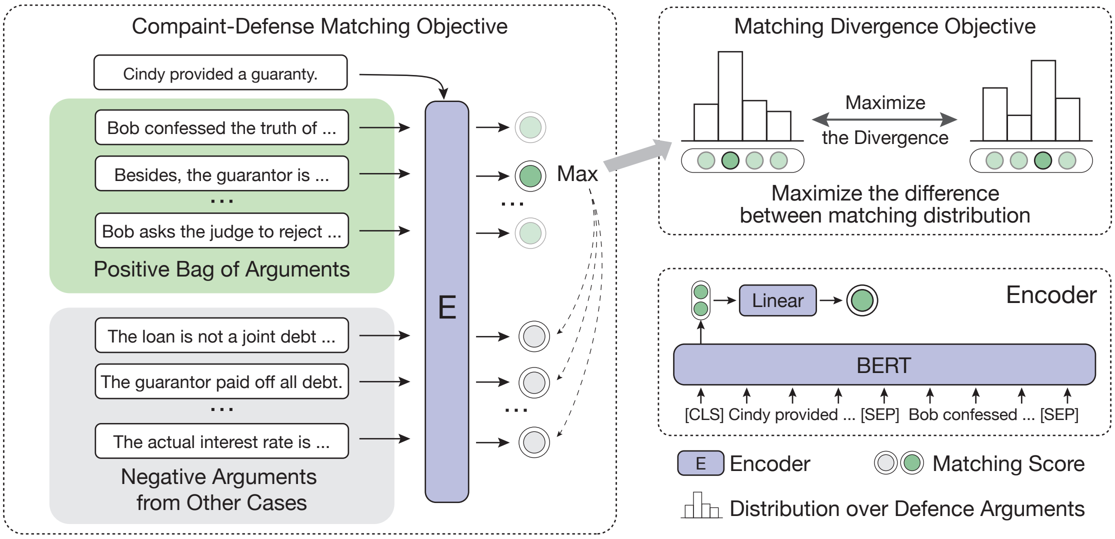
Compaint-Defense Matching Objective
Cindy provided a guaranty.
Bob confessed the truth of ...
Besides, the guarantor is ...
...
Bob asks the judge to reject ...
Positive Bag of Arguments
The loan is not a joint debt ...
The guarantor paid off all debt.
...
The actual interest rate is ...
Negative Arguments from Other Cases
E
Max
Matching Divergence Objective
Maximize
the Divergence
Maximize the difference between matching distribution
Linear
Encoder
BERT
[CLS] Cindy provided ... [SEP] Bob confessed ... [SEP]
E Encoder
Matching Score
Distribution over Defence Arguments

Figure 2: The illustration of the proposed model. (1) The complaint-defense matching objective requires the model to assign high matching scores to related arguments from the same documents with arguments from other cases as negative candidates. (2) The matching divergence objective requires the model to match different defense arguments for different complaint arguments.

this input sequence and generates a set of hidden representations, where the final hidden representation of the `[CLS]` token, denoted as $\mathbf{h}_c$ captures the contextual information of both arguments in the pair (q, p). To determine the matching score between the two arguments q and p, this `[CLS]` representation $\mathbf{h}_c$ is passed through a feed-forward neural network, and the output is used as the matching score, quantifying the degree of correlation or contrast between the two arguments in question.

$$s(q,p) = \mathbf{W}\mathbf{h}_c + b, \tag{1}$$

where $\mathbf{W}$ and b are trainable parameters. A higher score indicates a higher probability that the two arguments discuss the same topic.

3.4. Pre-training Objectives

Complaint-Defense Matching Objective The target of argument-pair extraction is to match the arguments about the same topic. We find that each complaint argument is supposed to be responded and it is also intuitive that argument pairs must come from complaints and defenses of the same cases, which can provide weak supervision for the argument-pair extraction task. Inspired by at-least-one strategy in distantly supervised relation extraction (Zeng et al., 2014), after given a complaint argument, we treat the corresponding defense as a positive bag, in which there is at least one positive interactive argument. We greedily match the arguments by applying the maximum operation to select the arguments with the same topics. Specifically, given the complaint argument $q \in c$, we calculate the positive matching score as:

$$s^+(q) = \max_{p_j \in d^+} s(q, p_j), \tag{2}$$

where d^+ is the corresponding defense from the same case as c. Intuitively, the defense argument with the highest matching score is treated as the positive instance in further contrastive learning.

To enforce the model to assign high scores to related argument pairs, we train the model with the contrastive learning mechanism. We randomly sample the negative defense arguments from other cases and require the model to correctly match the complaint arguments and defense arguments from the same cases. Specifically, given a complaint argument $q \in c$ and the related defense d^+ recorded in the same case, we first randomly select N irrelevant defense arguments from other cases as the negative candidates $\{p_1^-, ..., p_N^-\}$. We define the complaint-defense matching objective as:

$$\mathcal{L}_1 = -\log \frac{\exp(s^+(q))}{\exp(s^+(q)) + \sum_{k=1}^{N} \exp(s(q, p_k^-))}. \tag{3}$$

This objective encourages the related argument pairs to have a higher score than unrelated ones. Notably, prior works prove that the hard negatives are needed to facilitate better training (Robinson et al., 2021; Kalantidis et al., 2020). Therefore, we sample negatives from cases of the same cause, which usually involve similar events and disputes. The hard negatives can help the model to compare the argument pairs in detail. Notably, though arguments from the same cause may be similar, they are usually different in case details. We also remove common arguments with no discussion of specific facts, such as “have no opposite opinions" and “request the judge to deny all appeals". Therefore, the false negative problem can be avoided.

Matching Divergence Objective The complaint-defense matching objective utilizes supervision from the level of complaint-defense pairs, which enables the model to be trained on the unlabeled dataset. However, it may lead the model to learn trivial shortcuts. As mentioned before, we sample arguments from other cases as negative instances. Thus, the model can be optimized to give high scores to all defense arguments from the same case, which may result in only identifying entities, such as names, locations, and times, not the topics and details in the arguments.

To tackle this issue, we propose the matching divergence objective which is based on the observation that different complaint arguments tend to match different defense arguments. Therefore, the model should not give high scores to all argument pairs from the same cases. Specifically, the matching divergence objective encourages the matching distribution of complaint arguments over defense arguments to be diverse. For each training iteration, we first sample two complaint arguments, q_i and q_l, from the same document. We then calculate matching distribution over the defense arguments, $d = \{p_1, ..., p_m\}$, with a softmax normalization:

$$P_q(p_j) = \text{softmax}\left(s(q, p_j)\right), (p_j \in d, q = q_i, q_l)\,. \quad (4)$$

We use Jensen–Shannon divergence to measure the difference between two matching distributions and calculate matching divergence objective as: $\mathcal{L}_2 = 1 - \text{JS}(P_{q_i} || P_{q_l})$. This objective encourages two distributions P_{q_i} and P_{q_l} to be different, i.e., different complaint arguments should match different defense arguments.

To summarize, for each iteration, we sample two complaint arguments from the same document. The final loss is calculated as the sum of the contrastive matching loss and the divergence loss: $\mathcal{L}_c = \mathcal{L}_1 + \lambda \mathcal{L}_2$, where λ is a hyper-parameter. As the relation between complaint arguments and defense arguments is a two-way relation, both two assumptions are also valid for defense arguments. That is to say, for each defense argument, we can find at least one corresponding complaint argument, and the matching relationship over complaint arguments should be different. Therefore, just as the definition of $\mathcal{L}_c$, we can define the loss function for defenses $\mathcal{L}_d$ as the sum of the two objectives. The final pre-training loss function is $\mathcal{L} = \mathcal{L}_c + \mathcal{L}_d$.

3.5. Downstream Evaluation

We evaluate our model with the human-annotated data under two settings: unsupervised setting and supervised setting. The unsupervised setting does not require extra labeled data for training, while the supervised setting can leverage human-annotated data to further improve the model performance.

For *unsupervised setting*, we directly employ the pre-trained model to extract the argument pairs based on the matching scores in Eq. 1. The defense arguments assigned with the highest matching score are selected as the positive answers.

For *supervised setting*, we further utilize annotated data to fine-tune the pre-trained model. Similar to the complaint-defense matching objective, we require the model to distinguish the related argument, and the objective is defined as:

$$\mathcal{L}_f = -\text{log}\frac{\text{exp}(s(q, p^+))}{\text{exp}(s(q, p^+)) + \sum_i \text{exp}(s(q, p_i^-))}, \quad (5)$$

where p_i^- is the sampled negative defense argument. Different from the negative samples in the pre-training stage, the negative samples in the fine-tuning stage are provided by the human-annotated dataset, which comes from the same case as the complaint argument. It is more challenging to distinguish the positive argument from the human-annotated negative arguments. Therefore, fine-tuning can further facilitate better argument pair extraction and achieve higher accuracy.

4. Experiment

4.1. Dataset Construction

To evaluate the effectiveness of our model, we construct two legal argument-pair extraction datasets based on large-scale open-access legal documents published by the Chinese government[1]. We select cases from two representative causes with the most cases, including private lending (Lending) and contract dispute (Contract). Both two types of cases are common and important in our daily life.

To improve data quality, we only preserve the documents from the intermediate courts and discard the documents from grass-roots courts. As the legal documents are well-formatted, we can apply hand-crafted regular expressions to divide the documents into several parts. Firstly, for each case, we divide the document into five parts: the information about two parties, the complaint, the defense, the court views, and the judgment results, where only the complaints and defenses are kept. We then delete the common arguments which contain the words “having no opposite opinions" or “request the judge to deny all appeal". These common arguments do not contain specific details or information, and matching them cannot significantly benefit downstream tasks. The deletion is also conducted via regular expressions. Notably, the data

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

preprocessing is conducted without human annotation, and the required regular expressions are also low-costs. Therefore, we can easily conduct large-scale self-supervised pre-training.

We first collect unlabeled cases to form the pre-training datasets, and then select thousands of cases to perform human annotation for fine-tuning and evaluation. To ensure high data quality, we require that all annotators are in a law-related profession or pursuing a law degree. Besides, before carrying out annotation, annotators have to go through several hours of training for labeling. Besides, during mannual annotation, we adopt honeypot data, where the correct annotations are already known. When annotators mistakenly annotate honeypot data, we will pause their annotation to keep them focused during the annotation process. We adopt a two-stage annotation process for dataset construction. Firstly, each case is required to be annotated twice independently. We employ Cohen's Kappa (Cohen, 1960) to measure the inter-agreements between two annotators. The Kappa coefficients are 44% and 46% for private lending cases and contract dispute cases. Secondly, we require an experienced annotator to give the final results based on the results given in the first step. Only the annotators with high consistency in the first step are allowed to participate in the second step. To evaluate data quality, we randomly sample 10% cases from two datasets to check the data quality. The estimated precision of the annotation is 95.26%. The estimated recall of the annotation is 91.70%. From the results, we can observe that though the legal argument-pair extraction task is challenging, our datasets are high-quality and can serve as a good benchmark for future research.

The annotated data are randomly split into the training set, validation set, and test set. The detailed statistics are listed in Table 1. Following previous work (Yuan et al., 2021a), for each complaint argument, we give 5 defense arguments as candidates where only one is the positive related argument. Our datasets consist of thousands of complicated legal cases, and are the largest legal argument-pair extraction dataset. Note that CAIL2020-Argmine dataset (Yuan et al., 2021a) does not provide information on the distribution of legal causes, which makes it difficult to obtain appropriate pre-training data. Therefore we do not evaluate our model on CAIL2020-Argmine.

4.2. Baseline Models

Unspervised Baselines Relatively few studies have ventured into the realm of unsupervised argument-pair extraction. Given that interactive argument pairs are thematically coherent, we select state-of-the-art sentence representation learning methods as our baseline models for comparison. These models are: 1) **AvgEmb** calculates the representation of an argument as the mean of its token representations. The matching score between two arguments is defined as the cosine similarity between their average embeddings. 2) **IS-BERT** (Zhang et al., 2020) maximizes mutual information between local and global sentence representations. In other words, it seeks to ensure that individual tokens and the sentence as a whole share high information content, thereby creating robust sentence representations suitable for matching. 3) **SimCSE** (Gao et al., 2021) uses a contrastive learning approach and crafts positive examples by applying dropout twice on the same sentence. The model then minimizes the distance between these perturbed versions in the representation space while maximizing the distance to other unrelated sentences. Both **IS-BERT** and **SimCSE** are pre-trained on the same corpus that we use for our own method, ensuring a fair comparison.

Dataset	pre-train	train	valid	test
		Lending		
# Case	20k	1.2k	400	400
# Argument	194k	11.7k	3.9k	3.9k
# Arg-Pair	–	11.9k	4.4k	4.8k
		Contract		
# Case	20k	1.2k	400	400
# Argument	201k	12.4k	4.2k	4.1k
# Arg-Pair	–	11.3k	4.5k	4.8k

Table 1: The detailed statistics of datasets. Here, # Case, # Argument, and # Arg-Pair denote the number of cases, arguments, and annotated positive argument-pairs.

Spervised Baselines For evaluation in supervised setting, we select a variety of state-of-the-art models that specialize either in sentence representation learning or in argument-pair extraction:

1) **SBERT** (Reimers and Gurevych, 2019) utilizes a siamese architecture and utilizes a triplet distance objective to optimize the model. 2) **SimCSE$_{\text{Sup}}$** (Gao et al., 2021) is the supervised version of SimCSE. It uses labeled positive examples to carry out contrastive learning, aiming to generate highly discriminative sentence representations. 3) **BERT-Pair** (Devlin et al., 2019) concatenates the arguments from both the complaint and defense and feeds them into a BERT encoder. It serves as a straightforward application of BERT to our problem and can be viewed as our model without additional pre-training steps. 4) **DARL** (Ji et al., 2021) employs a discrete auto-encoder to capture various facets of information within arguments. The original version uses a GRU as the encoder, but for a fair comparison, we replace it with BERT.

Notably, to make a fair comparison, we em-

Model	Lending		Contract	
	valid	test	valid	test
Unsupervised Setting				
AvgEmb	36.7	36.5	43.1	43.3
IS-BERT	40.9	41.0	51.8	52.6
SimCSE	42.9	41.7	51.8	51.9
Ours	**48.4**	**46.4**	**53.4**	**54.6**
Supervised Setting				
SBERT	54.6	54.0	60.6	59.8
SimCSE$_{\text{Sup}}$	57.5	58.0	61.2	61.1
BERT-Pair	62.0	62.0	65.3	65.5
DARL	58.4	57.6	65.8	65.2
Ours$_{\text{Sup}}$	**65.6**	**64.0**	**68.0**	**68.3**

Table 2: The performance of the baseline models and our proposed model (Accuracy).

ploy BERT pre-trained on Chinese civil case documents (Zhong et al., 2019) as the encoder for all models. As the cases in our datasets are part of this pre-training corpus, further pre-training on the same data does not offer additional benefits, which is confirmed in our ablation studies. This setup guarantees that each model is given an equal opportunity to showcase its capabilities in LAE.

4.3. Implementation Details

Our experimental setup is carefully designed to ensure reproducibility and to provide a fair comparison across models. Here are the key details: During the pre-training stage, we set the batch size as 128 implemented through gradient accumulation. For the pre-training learning rate and λ in the objective function, we apply the grid search strategy to select the best hyper-parameters. We select the pre-training learning rate from $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}\}$, with the best-performing rate set to 1×10^{-4}. We select the hyper-parameter λ for the objective function from $\{0.1, 0.2, 0.4, 0.5, 1\}$, with 0.1 being the chosen value for the Lending dataset and 0.2 for the Contract dataset.

4.4. Main Results

The comparison results between our methods and baseline models are shown in Table 2. From the results we can observe that: 1) Our proposed model can significantly outperform the competitive baseline models by a significant margin under both the unsupervised setting and the supervised setting. It proves that our proposed model can effectively capture the semantic information, and thus achieve a performance improvement. 2) Focusing on the unsupervised setting, it's worth noting that both IS-BERT and SimCSE also take advantage of unlabeled data for their pre-training phases. Despite this, our model still demonstrates superior performance. This implies that the unsupervised learning objectives we introduce are particularly effective at enabling the model to recognize and capture topical features. This capability is beneficial for the downstream task of argument-pair extraction. 3) While our unsupervised model does achieve a considerable gain in performance, there is still a discernible performance delta when compared to supervised models. This observation highlights the value of high-quality, human-annotated data in refining the model's capabilities. Additionally, it serves as a call to action for researchers to direct more effort into enhancing the performance of unsupervised models in the area of argument-pair extraction. This is an avenue we earmark for future research endeavors. 4) Another observation is the architectural choices in encoding argument pairs. Both BERT-Pair and our model utilize a single BERT encoder to process given argument pairs, contrasting with other models that employ a Siamese architecture. Our results substantiate that the joint modeling of given sentences is essential for effective LAE.

Model	Lending		Contract	
	valid	test	valid	test
Unsupervised Setting				
Ours	**48.4**	**46.4**	**53.4**	**54.6**
w/o divergence	48.1	45.4	52.1	52.4
Supervised Setting				
Ours$_{\text{Sup}}$	**65.6**	**64.0**	**68.0**	**68.3**
w/o divergence	65.3	63.6	67.8	67.5
w/o pre-training	62.0	62.0	65.3	65.5
w/ MLM	62.5	62.3	65.2	65.0

Table 3: Results of the ablation study (Accuracy).

In summary, our model can effectively leverage large-scale unlabeled data to conduct pre-training, and improve the semantic matching ability. We argue that we should attach more importance to the utilization of unsupervised data for future research.

4.5. Ablation Study

To dive into the effectiveness of our proposed model, we conduct an ablation study and the results are shown in Table 3. In this study, we manipulate two key factors: the matching divergence objective and the pre-training process. Specifically, the term **w/o divergence** designates a model that is pre-trained without the utilization of the matching divergence objective. On the other hand, **w/o pre-training** denotes a model that is fine-tuned without any prior pre-training. Besides, to further verify the effectiveness of our pre-training tasks, we also present the results of the model with further masked language model (MLM) pre-training on the unsupervised corpus (**w/ MLM**). We can observe that:

	Lending		Contract	
Strategy	valid	test	valid	test
	Unsupervised Setting			
SameType	**48.4**	**46.4**	**53.4**	**54.6**
Random	45.0	41.7	49.1	48.7
Overlap	47.4	44.9	52.4	53.6
	Supervised Setting			
SameType	**65.6**	**64.0**	68.0	**68.3**
Random	64.5	62.9	66.8	66.3
Overlap	64.7	63.6	**69.4**	**68.3**

Table 4: Results for three negative sampling strategies (Accuracy).

1) The matching divergence objective and the pre-training process contribute to the main model. Removing either one of these elements results in a noticeable drop in performance metrics. 2) Strikingly, the absence of the matching divergence objective yields a significant performance decline in the unsupervised setting—amounting to a drop of 1.0 and 2.2 points across two test datasets. However, the impact on the supervised setting is markedly less severe, registering a decline of only 0.4 and 0.8 points respectively across the test datasets. This suggests that the matching divergence objective functions as an effective regularization mechanism in unsupervised learning. In supervised settings, this objective seems less critical since high-quality, human-annotated data already fulfill the regularization requirements. Therefore, its omission does not significantly hinder performance in a supervised environment. 3) Interestingly, further pre-training using the MLM objective fails to yield consistent improvements. Given that the backbone BERT architecture has already undergone pre-training using legal documents, additional in-domain MLM pre-training does not contribute to performance enhancement. This contrasts with our argument-specific pre-training objectives, which demonstrate effectiveness in downstream LAE performance.

4.6. Effects of Negative Sampling

Negative sampling is a pivotal component in the realm of contrastive learning, as established by existing literature (Kalantidis et al., 2020; Robinson et al., 2021). In this section, we specifically examine the influence of different negative sampling strategies on the performance of our model. As mentioned before, our model is trained with negative samples sourced from cases that share the same cause to facilitate better training, referred to as the **SameType** strategy. Additionally, we explore two alternative strategies: 1) **Random**, where negative samples are randomly drawn from irrelevant cases, and 2) **Overlap**, where negatives are selected based on minimal word overlap with the query argument and are sourced from the same case document. The impact of these sampling strategies is empirically captured in Table 4. Our observations indicate: 1) Utilizing randomly selected negatives during the pre-training phase results in a substantial decline in performance across both supervised and unsupervised settings. This suggests that the nature of the negative samples matters significantly, affirming that hard negatives are indispensable for effective contrastive pre-training. 2) Contrary to expectations, the Overlap strategy does not outperform the SameType strategy. The reason for this is that the Overlap strategy inadvertently introduces false negatives into the model's training, thereby negatively impacting its performance.

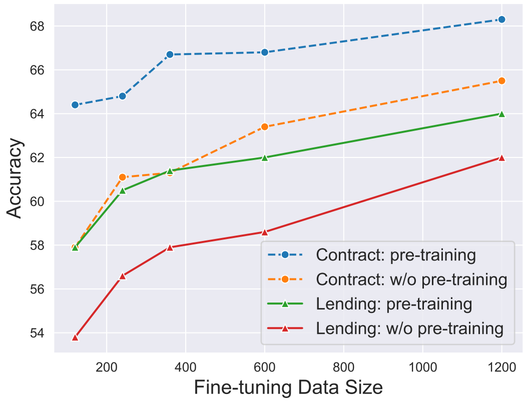

Figure 3: Supervised performance (accuracy) with different fine-tuning data size. Best viewed in color.

These findings underscore the necessity of carefully crafting negative sampling strategies. Specifically, it indicates that the selection of appropriately challenging negatives has the potential to further boost the performance of LAE.

4.7. Effects of Supervised Data Size

We argue that our model can achieve superior results with limited supervised data. To substantiate this claim, we investigate how performance varies with different sizes of fine-tuning datasets. We compare the performance of our pre-trained model with that of a model that has not undergone any pre-training. The results are shown in Figure 3. Our model demonstrates remarkable performance in scenarios where supervised data is sparse. Specifically, it is capable of delivering competitive or superior results even when only 50% of labeled data is available for fine-tuning. We also notice a trend of performance improvement correlating with an increase in the size of supervised dataset. This suggests that there is room for enhancing the model's performance by adding more annotated data.

These findings validate our hypothesis about the effectiveness of our model in low-resource settings, allowing for high performance even when labeled data is limited. Therefore, the model serves as a practical solution in real-world scenarios where high-quality annotated data is often costly to acquire, while still offering the potential for further performance gains with additional data.

5. Conclusion and Future Work

In this paper, we focus on the LAE task, aiming to address the challenge of data sparsity that often plagues this domain. To this end, we introduce a contrastive pre-training method that leverages the wealth of unsupervised legal documents available. By doing so, we enable our model to learn and capture argumentative knowledge effectively. We incorporate two self-supervised objectives that are formulated based on intuitive observations. Our experiments indicate that these objectives contribute to a significant boost in performance. In the future, we will explore utilizing the unlabeled data more effectively, and enable the unsupervised model to reach comparable results with supervised models.

Ethical Consideration

In this section, we delve into ethical considerations that arise from the deployment of our proposed model for argument-piring extraction from legal documents.

1) **Potential for Misuse:** The primary intent behind our model is to augment the process of argument-pair extraction in the legal domain, thereby aiding downstream tasks like case analysis and dispute focus summarization. This has considerable value, particularly in legal consultation systems. However, the model is susceptible to misuse, especially if it is perceived as a full-fledged replacement for human legal practitioners. We strongly contend that our model should serve merely as an assistive tool for legal experts and not replace them. The final judgement in legal cases should unequivocally rest with duly appointed judges to mitigate the risk of model misuse.

2) **Data Sourcing and Annotation:** The dataset leveraged for this research has been sourced from publicly accessible documents released by the Supreme People's Court of China. It's important to clarify that the data is not confidential and is freely available to the general public. Regarding the annotation process, we started by manually annotating a subset of examples to gauge the workload. Subsequently, we compensated annotators in accordance with local wage standards.

Acknowledgement

This work is supported by Quan Cheng Laboratory (Grant No. QCLZD202301). This work is also supported by Huawei MindSpore team.

In the preparation and discussion of the project, Chaojun Xiao and Yuan Yao designed the model architectures. Yutao Sun and Wenbin Zhang wrote the code and conducted the experiments. Chaojun Xiao and Yutao Sun wrote the initial draft. Yuan Yao, Xu Han, and Zhiyuan Liu significantly edited and improved the paper. Maosong Sun provided valuable advice to the research.

References

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021a. A neural transition-based model for argumentation mining. In *Proceedings of ACL-IJCNLP*, pages 6354–6364. Association for Computational Linguistics.

Jianzhu Bao, Bin Liang, Jingyi Sun, Yice Zhang, Min Yang, and Ruifeng Xu. 2021b. Argument pair extraction with mutual guidance and intersentence relation graph. In *Proceedings of EMNLP*, pages 3923–3934. Association for Computational Linguistics.

Jianzhu Bao, Jingyi Sun, Qinglin Zhu, and Ruifeng Xu. 2022. Have my arguments been replied to? argument pair extraction as machine reading comprehension. In *Proceedings of ACL*, pages 29–35. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of EMNLP*, pages 39–49. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing

Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of ACL*, pages 4317–4323. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: "preparing the muppets for court'". In *Proceedings of EMNLP: Findings*, volume EMNLP 2020 of *Findings of ACL*, pages 2898–2904. Association for Computational Linguistics.

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of COLING*, pages 1561–1571. International Committee on Computational Linguistics.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of EMNLP*, pages 7000–7011. Association for Computational Linguistics.

Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. 2021. Argument pair extraction via attention-guided multi-layer multi-cross encoding. In *Proceedings of ACL/IJCNLP*, pages 6341–6353. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*, abs/2306.16092.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. 2019. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In *Proceedings of CIKM*, pages 1361–1370. ACM.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910. Association for Computational Linguistics.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in CJEU decisions on fiscal state aid. In *Proceedings of ArgMining@COLING*, pages 143–157. International Conference on Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2022. Mining legal arguments in court decisions. *CoRR*, abs/2208.06178.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. In *Proceedings of NeurIPS*.

Lu Ji, Zhongyu Wei, Jing Li, Qi Zhang, and Xuanjing Huang. 2021. Discrete argument representation learning for interactive argument pair identification. In *Proceedings of NAACL-HLT*, pages 5467–5478. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard H. Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Trans. Assoc. Comput. Linguistics*, 9:721–739.

Yannis Kalantidis, Mert Bülent Sariyildiz, Noé Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. In *Proceedings of NeurIPS*.

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of ACL*, pages 7067–7072. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2021. Stance detection: A survey. *ACM Comput. Surv.*, 53(1):12:1–12:37.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. SAILER: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of SIGIR*, pages 1035–1044. ACM.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: A legal case retrieval dataset for chinese law system. In *Proceedings of SIGIR*, pages 2342–2348. ACM.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn A. Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of NAACL*, pages 430–440. The Association for Computational Linguistics.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of ICAIL*, pages 225–230. ACM.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of ICAIL*, pages 98–107. ACM.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of NAACL*, pages 1384–1394. The Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3980–3990. Association for Computational Linguistics.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *Proceedings of ICLR*. OpenReview.net.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. BERT-PLI: modeling paragraph-level interactions for legal case retrieval. In *Proceedings of IJCAI*, pages 3501–3507. ijcai.org.

Lida Shi, Fausto Giunchiglia, Rui Song, Daqian Shi, Tongtong Liu, Xiaolei Diao, and Hao Xu. 2022. A simple contrastive learning framework for interactive argument pair identification via argument-context extraction. In *Proceedings of EMNLP*, pages 10027–10039. Association for Computational Linguistics.

Reid Swanson, Brian Ecker, and Marilyn A. Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of SIGDIAL*, pages 217–226. The Association for Computer Linguistics.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of AAAI*, pages 9048–9056. AAAI Press.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of ACL-IJCNLP*, pages 1338–1352. Association for Computational Linguistics.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of EMNLP*, pages 763–780. Association for Computational Linguistics.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Chaojun Xiao, Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2023. Legal knowledge representation learning. In *Representation Learning for Natural Language Processing*, pages 401–432. Springer.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale chinese legal event detection dataset. In *Proceedings of ACL: Findings*, pages 183–201. Association for Computational Linguistics.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of NAACL-HLT*, pages 1854–1864. Association for Computational Linguistics.

Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2022. Explainable legal case matching via inverse optimal transport-based rationale extraction. In *Proceedings of SIGIR*, pages 657–668. ACM.

Jian Yuan, Zhongyu Wei, Yixu Gao, Wei Chen, Yun Song, Donghua Zhao, Jinglei Ma, Zhen Hu, Shaokun Zou, Donghai Li, and Xuanjing Huang. 2021a. Overview of smp-cail2020-argmine: The interactive argument-pair extraction in judgement document challenge. *Data Intell.*, 3(2):287–307.

Jian Yuan, Zhongyu Wei, Donghua Zhao, Qi Zhang, and Changjian Jiang. 2021b. Leveraging argumentation knowledge graph for interactive argument pair identification. In *Proceedings of ACL: Findings*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2310–2319. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, pages 2335–2344. ACL.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of EMNLP*, pages 1601–1610. Association for Computational Linguistics.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of EMNLP*, pages 3540–3549. Association for Computational Linguistics.

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 ACL*, pages 5218–5230. Association for Computational Linguistics.

Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Open Chinese language pre-trained model zoo. *Technical report*.