Explainable Legal Case Matching via Inverse Optimal Transport-based Rationale Extraction

Weijie Yu
School of Information
Renmin University of China
yuweijie@ruc.edu.cn

Zhongxiang Sun, Jun Xu*
Gaoling School of Artificial Intelligence
Renmin University of China
jeryi.sunzx01@gmail.com,junxu@ruc.edu.cn

Zhenhua Dong
Noah's Ark Lab, Huawei
dongzhenhua@huawei.com

Xu Chen
Gaoling School of Artificial Intelligence
Renmin University of China
xu.chen@ruc.edu.cn

Hongteng Xu
Gaoling School of Artificial Intelligence
Renmin University of China
hongtengxu@ruc.edu.cn

Ji-Rong Wen
Gaoling School of Artificial Intelligence
Renmin University of China
jrwen@ruc.edu.cn

ABSTRACT

As an essential operation of legal retrieval, legal case matching plays a central role in intelligent legal systems. This task has a high demand on the explainability of matching results because of its critical impacts on downstream applications — the matched legal cases may provide supportive evidence for the judgments of target cases and thus influence the fairness and justice of legal decisions. Focusing on this challenging task, we propose a novel and explainable method, namely *IOT-Match*, with the help of computational optimal transport, which formulates the legal case matching problem as an inverse optimal transport (IOT) problem. Different from most existing methods, which merely focus on the sentence-level semantic similarity between legal cases, our IOT-Match learns to extract rationales from paired legal cases based on both semantics and legal characteristics of their sentences. The extracted rationales are further applied to generate faithful explanations and conduct matching. Moreover, the proposed IOT-Match is robust to the alignment label insufficiency issue commonly in practical legal case matching tasks, which is suitable for both supervised and semi-supervised learning paradigms. To demonstrate the superiority of our IOT-Match method and construct a benchmark of explainable legal case matching task, we not only extend the well-known Challenge of AI in Law (CAIL) dataset but also build a new Explainable Legal cAse Matching (ELAM) dataset, which contains lots of legal cases with detailed and explainable annotations. Experiments on these two datasets show that our IOT-Match outperforms state-of-the-art methods consistently on matching prediction, rationale extraction, and explanation generation.

*Jun Xu is the corresponding author. Work partially done at Beijing Key Laboratory of Big Data Management and Analysis Methods.

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531974

CCS CONCEPTS

• **Applied computing** → **Law**; • **Information systems** → **Information extraction**.

KEYWORDS

Legal retrieval, Explainable matching

ACM Reference Format:
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 the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3531974

1 INTRODUCTION

Legal case matching aims at identifying relations between paired legal cases, which is a key task of legal retrieval. This task has a high demand on the explainability of matching results because of its critical impacts on legal justice. In particular, the matching results may indicate significant evidence or information, which influences the incentives of decision-makers in the common law system and provides the basis for legal reasoning in the civil law system.

To achieve this aim, many efforts have been made, including the early attempts that are based on rule-based strategies [5, 39, 52] and the recent learning-based methods like the Precedent Citation Network (PCNet) [25] and BERT-based methods [41, 44]. Although these methods have achieved encouraging performance, they often suffer from the following challenges on providing plausible and faithful explanations associated with matching results. Firstly, legal cases are long-form documents with complicated contents in general, in which only the rationales representing the legal characteristics should support matching results and their explanations. However, existing methods tend to overlook the striking different roles between the rationales and other sentences [11, 41, 44]. Secondly, ideal explanations of legal case matching results are expected to offer reasons for one side and to rebut arguments for the other side [3], but existing methods often fail to distinguish the pro rationales and con rationales that respectively support the matching and

mismatching decisions [30, 34, 40]. Moreover, the ground-truth labels for explanations (e.g., aligned rationales across different cases) are expensive, sparse, and usually biased (e.g., only limited number of positive pairs are labeled correctly while lots of false negative pairs exist). As a result, learning based on such labeled data often leads to sub-optimal matching results and unreliable explanations.

Facing the above challenges, in this paper, we propose a novel inverse optimal transport [20, 27] (IOT)-based model, called IOT-Match, to extract rationales for explainable legal cases matching. As illustrated in Figure 1, our IOT-Match formulates the extraction and alignment of pro and con rationales as an optimal transport (OT) problem, in which the identified rationales and their alignments are derived from the transport plan of the OT solution. The optimal transport is guided by a learnable affinity matrix that reflects both semantics and legal characteristic relations between cross-case sentences, and the affinity matrix is learned by the IOT process, which corresponds to solving a bi-level optimization problem. In this way, IOT-Match learns to extract the pro and con rationales directly. To apply the proposed model to real legal case matching applications and following the practices in [26, 54], the extracted rationales from the paired legal cases are then fed to a pre-trained language model to generate label-specific natural languages explanations that stand for the pro and con reasons of matching. For filtering out the noise sentences and weighing the pro and con reasons, the final matching results are made based on the extracted rationales and the generated label-specific explanations.

Besides proposing an explainable legal case matching method, we construct a new dataset called **E**xplanable **L**egal c**A**se **M**atching (ELAM). To be best of our knowledge, our ELAM is the first legal case matching dataset which provides matching labels for legal case pairs and detailed annotations, including rationales, alignments, and natural language-based explanations for matching labels.

In summary, our contributions include the following three folds: (1) We propose a novel model, namely IOT-Match, to extract rationales and generate natural language-based explanations for legal case matching. To the best of our knowledge, IOT-Match is the first explainable model for legal case matching. (2) We construct a new large-scale dataset ELAM which facilitates future research on explainable legal case matching. (3) Experimental results indicate IOT-Match not only achieves state-of-the-art performance in legal case matching but also produces plausible and faithful explanations for its matching prediction.

2 RELATED WORK

2.1 Legal Case Matching

Conventional legal case matching methods highly depend on expert knowledge [5], e.g., the decomposition of legal issues [52] and the ontological framework of the problem [39]. In recent years, learning-based legal case matching strategy has shown advantages in exploring the semantics of legal cases, which can be roughly categorized into network-based methods [7, 25, 32, 33] and text-based methods [41, 44]. The network-based methods construct a Precedent Citation Network (PCNet), in which the vertices are legal cases and directed edges indicate the citations of source cases used by target cases. Based on PCNet, Kumar et al. [25] used the Jaccard similarity index between the sets of precedent citations to infer the

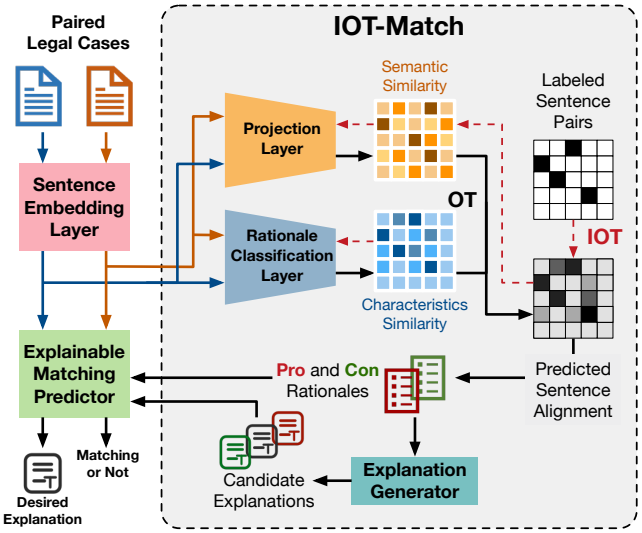


Figure 1: The architecture of our model IOT-Match. Note that the red dotted arrows indicate the back-propagation achieved by inverse optimal transport, which are used only in the training phase.

similarity of two legal cases. Minocha et al. [32] used whether the sets of precedent citations occurs in the same cluster to measures to what extent the two cases are similar. Bhattacharya et al. [7] proposed Hier-SPCNet to capture all domain information inherent in both statutes and precedents. The text-based methods rely on the textual content of the cases and measure the similarity of two legal cases based on their semantics. Shao et al. [41] proposed BERT-PLI to break a case into paragraphs and model the interactions between the paragraphs. It first adopted BERT to encode each paragraph in two legal cases, then applied max-pooling to capture their matching signal, and finally, used a recurrent neural network (RNN) with an attention mechanism to predict their matching score. Similarly, Bhattacharya et al. [8] proposed to segment two legal cases into paragraphs and aggregate the paragraph-level similarity. Inspired by the success of pre-trained language models in the generic domain, Xiao et al. [44] pre-trained a Longformer-based language model with tens of millions of criminal and civil case documents. Although these studies effectively improve the performance, they often have difficulties on explaining their predictions, which limits their practical applications [9].

2.2 Explainable AI in the Legal Domain

Recently, researchers have made some efforts to achieve explainable AI models in various applications of the legal domain [19]. In the task of legal judgment prediction, Ye et al. [48] formalized the court view generation problem as a label-conditioned Seq2Seq task and generated court views based on fact descriptions and charges. Jiang et al. [23] proposed a neural based system to jointly extract readable rationales and elevate charge prediction accuracy by a rationale augmentation mechanism. Liu et al. [28] proposed the Joint Prediction and Generation Model (JPGM) to predict charges and court views. JPGM generated charge-discriminative information and used

a coarse-to-fine classifier to effectively deal with the confusions within charges. In addition, JPGM explicitly modeled the interdependence between charges and court views. In the task of legal question answering, Zhong et al. [55] proposed to first detect elements of fact descriptions by iteratively asking questions about pre-defined charge-specific principles and then used the detected elements for prediction. Different from the above work, we focus on the legal case matching task, extracting rationales and generating natural language-based explanations to support matching results.

3 PROPOSED IOT-MATCH METHOD

3.1 Problem Statement

To conduct explainable legal case matching, we are given a set of labeled data tuples $\mathcal{D} = \{(X, Y, \mathbf{r}^X, \mathbf{r}^Y, \hat{\mathbf{A}}, z, e)\}$. For each tuple $(X, Y, \mathbf{r}^X, \mathbf{r}^Y, \hat{\mathbf{A}}, z, e)$ in the dataset, its elements include: 1) a pair of legal cases $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ represent the sets of source and target legal cases; 2) the rationale labels of the paired cases, denoted as $\mathbf{r}^X$ and $\mathbf{r}^Y$, respectively; 3) a binary alignment matrix $\hat{\mathbf{A}}$ indicating the semantic relation between rationales of X and rationales of Y; and 4) the matching label z and the set of sentences (denoted as e) explaining the reasons for z.

In practice, we represent each legal case as a set of sentence-level embeddings, i.e., $X = \{x_m\}_{m=1}^M$ and $Y = \{y_n\}_{n=1}^N$, where x_m (y_n) denotes the embedding of the m-th (n-th) sentence in X (Y). Typically, each embedding can be calculated by using the output of at the [CLS] token of a BERT model pre-trained on a Chinese legal case corpus. [1] The rationale labels are associated with the sentence embeddings, i.e., $\mathbf{r}^X = \{r_{x_m}\}_{m=1}^M$ and $\mathbf{r}^Y = \{r_{y_n}\}_{n=1}^N$, where the rationale label of a sentence s is designed following [31]:

$$r_s = \begin{cases} 0 & s \text{ is not a rationale,} \\ 1 & s \text{ is a key circumstance,} \\ 2 & s \text{ is a constitutive element of crime,} \\ 3 & s \text{ is a focus of disputes.} \end{cases} \tag{1}$$

The remaining elements, i.e., $\hat{\mathbf{A}} = [\hat{a}_{mn}] \in \{0, 1\}^{M \times N}$, $z \in \{0, 1, 2\}$, and e, are annotated manually, where

$$\hat{a}_{mn} = \begin{cases} 0 & r_{x_m} \neq r_{y_n}, \\ 1 & r_{x_m} = r_{y_n} \ \& \ x_m \cong y_n, \end{cases} \quad z = \begin{cases} 0 & \text{Mismatched } (X, Y), \\ 1 & \text{Partially matched,} \\ 2 & \text{Matched,} \end{cases} \tag{2}$$

where $x_m \cong y_n$ means the sentences corresponding to x_m and y_n are semantically-similar. $\hat{a}_{mn} = 1$ means aligned rationales while $\hat{a}_{mn} = 0$ means misaligned rationales. They provide pro and con evidence for matching prediction, respectively. Figure 2 shows an example of human labeled explainable legal case pair.

The proposed explainable legal case matching aims at learning the following three modules: 1) f_1 extracts aligned and misaligned rationales from the paired legal cases (**Sec. 3.3**); 2) f_2 generates candidate explanations based on the rationales extracted by f_1 (**Sec. 3.4**); and 3) f_3 predicts the final matching label based on the extracted rationales and generated explanations (**Sec. 3.5**).

[1] Corpus available at https://github.com/thunlp/OpenCLaP. Note that IOT-Match is applicable to the corpus of other languages with the corresponding embeddings.

3.2 The Principle of Our Method

As the key of the above three modules, f_1 is learned to fit the given alignment matrices $\hat{\mathbf{A}}$'s. As aforementioned, however, the alignment matrices are manually labeled, which are often very sparse and thus contain many false negative elements. To learn our model robustly from such imperfect data, we develop a novel learning paradigm from the viewpoint of optimal transport (OT). Essentially, optimal transport [36, 43] defines a distance between probability distributions, which has been widely used in many machine learning tasks, such as point cloud alignment [1, 2], graph matching [12, 45, 46], data clustering [10, 47], and sequence representations learning [13, 50, 51]. In our scenario, following existing studies [1, 12, 13, 46], given two sets of sentence embeddings (e.g., the X and Y mentioned above), we assume their empirical distributions to be uniform, i.e., $\boldsymbol{\mu} = \frac{1}{M}\mathbf{1}_M$ and $\boldsymbol{\nu} = \frac{1}{N}\mathbf{1}_N$, where $\mathbf{1}_D$ represents the D-dimensional all-one vector, and accordingly, compute the optimal transport distance between them in a discrete format [17]:

$$\begin{aligned} \mathbf{A}^* &= \arg\min_{\mathbf{A} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathbb{E}_{m,n \sim \mathbf{A}}[c(x_m, y_n)] \\ &= \arg\min_{\mathbf{A} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \sum_{m=1}^{M} \sum_{n=1}^{N} a_{mn} \cdot c(x_m, y_n), \end{aligned} \tag{3}$$

where $\mathbf{A} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\mathbf{A} \in \mathbb{R}_+^{M \times N} | \mathbf{A}\mathbf{1}_N = \boldsymbol{\mu}, \mathbf{A}^\top \mathbf{1}_M = \boldsymbol{\nu}\}$, which represents an arbitrary joint distribution of the sentences with marginals $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. $\mathbf{C} = [c(x_m, y_n)] \in \mathbb{R}^{M \times N}$ is a sentence-level affinity matrix, whose element $c(x_m, y_n)$ measures the discrepancy between the two sentences based on their embeddings. As shown in Eq. (3), the optimal transport distance actually corresponds to the minimum expectation of the sentence-level discrepancy, in which the optimal joint distribution $\mathbf{A}^*$ is called "optimal transport". The optimal transport matrix provides a soft matching between the sentence sets in a probabilistic way — the element of the optimal transport, *i.e.*, a^*_{mn}, indicates the probability of the coherency of x_m and y_n, which provides the evidence for their matching.

Note that Eq. (3) is a linear programming. To accelerate the computation of optimal transport, in practice we often introduce an entropic regularizer into Eq. (3), which leads to an entropic optimal transport problem [17]:

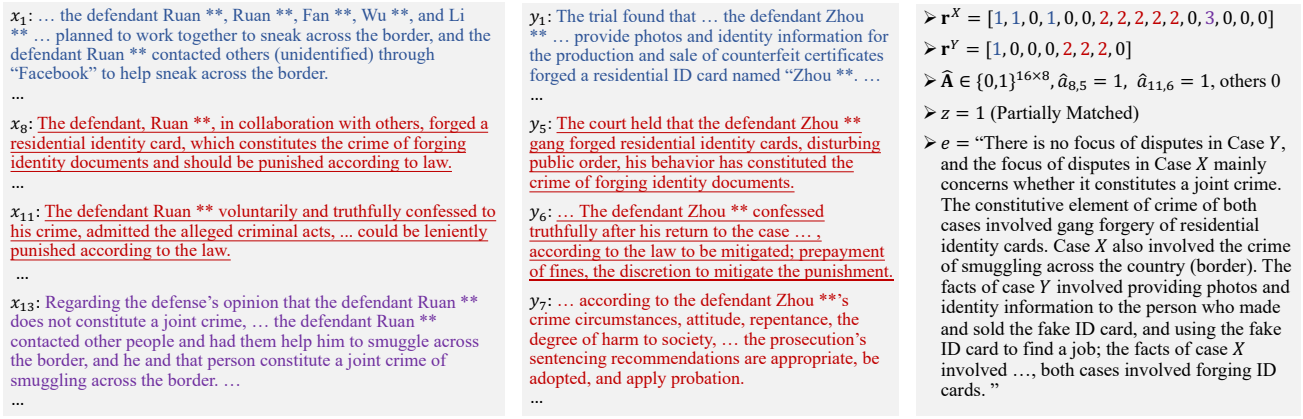
$$\mathbf{A}^* = \arg\min_{\mathbf{A} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{A}, \mathbf{C} \rangle + \gamma \langle \mathbf{A}, \log \mathbf{A} \rangle. \tag{4}$$

Here, $\langle \mathbf{A}, \mathbf{C} \rangle = \mathrm{Tr}(\mathbf{A}^\top \mathbf{C})$ represents the Frobenius dot-product, which represents the objective function of Eq. (3) in a matrix format. $\langle \mathbf{A}, \log \mathbf{A} \rangle = \sum_{m,n} a_{mn} \log a_{mn}$ is the proposed entropic regularizer. The entropic optimal transport in Eq. (4) is a strictly convex problem, which can be solved by the Sinkhorn scaling algorithm efficiently [17].

When learning the rationale extraction module f_1 in the above OT framework, the critical learning tasks become: 1) learning the affinity matrix $\mathbf{C}$ to optimize the guidance to the computation of the optimal transport $\mathbf{A}^*$; and 2) fitting the optimal transport $\mathbf{A}^*$ robustly to the manually-labeled (noisy) alignment matrix $\hat{\mathbf{A}}$. In this work, we solve these two tasks jointly by solving the following inverse optimal transport (IOT) problem.

3.3 IOT-based Rationale Extraction

According to the analysis above, we need to learn both the affinity matrix and the optimal transport based on the sentence embeddings



(a) case X (16 sentences) (b) case Y (8 sentences) (c) human annotations: $\mathbf{r}^X, \mathbf{r}^Y, \hat{\mathbf{A}}, z$, and e

Figure 2: A labeled legal case pair (translated from Chinese). Blue, red, and purple denote rationale labels of $r_s = 1, 2$ and 3, respectively. The underlined rationales are aligned. Note some sentences are omitted for the convenience of representation.

(X and Y) and their annotated alignment matrix $\hat{\mathbf{A}}$, which leads to a so-called inverse optimal transport (IOT) problem [20, 27]:

$$\begin{aligned} &\mathbf{C}^* = \arg\min_{\mathbf{C}\in\mathbb{R}^{M\times N}} \mathrm{KL}(\hat{\mathbf{A}}\|\mathbf{A}^*(\mathbf{C})), \\ &s.t.\ \mathbf{A}^*(\mathbf{C}) = \arg\min_{\mathbf{A}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{A}, \mathbf{C}\rangle + \gamma\langle \mathbf{A}, \log \mathbf{A}\rangle. \end{aligned} \tag{5}$$

This problem is a typical bi-level optimization problem, in which the affinity matrix $\mathbf{C}$ is the upper-level variable while the optimal transport $\mathbf{A}$ is the lower-level variable. The upper-level problem minimizes the KL divergence between $\hat{\mathbf{A}}$ and $\mathbf{A}^*$, i.e., $\mathrm{KL}(\hat{\mathbf{A}}\|\mathbf{A}^*) = \sum_{m,n} \hat{a}_{mn} \log \frac{a^*_{mn}}{\hat{a}_{mn}}$, which corresponds to the cross-entropy loss. The optimal transport $\mathbf{A}^*$ is a function of the affinity matrix, *i.e.*, $\mathbf{A}^*(\mathbf{C})$, whose optimization corresponds to the lower-level problem given $\mathbf{C}$.

Solving the IOT problem in Eq. (5) provides us a robust method to learn the rationale extraction module. Specifically, on the one hand, the upper-level problem fits the optimal transport to the limited and noisy alignment matrix under the constraint provided by the lower-level optimal transport problem, which suppresses the risk of over-fitting greatly. On the other hand, the lower-level problem provides us with an optimal transport matrix to indicate the aligned rationales, which is determined by the optimized affinity matrix and thus reveals sentence-level similarity between the paired legal cases. As a result, the optimal transport $\mathbf{A}^*$ derived from the optimal affinity matrix $\mathbf{C}^*$ represents the global alignment between rationales of a legal case pair. Accordingly, we can extract pro (aligned) and con (misaligned) rationales by setting a threshold τ, i.e., x_m and y_n are selected as pro rationales if $a^*_{mn} \geq \tau$, otherwise, are selected as con rationales.

Note that because the lower-level problem is strictly convex, this IOT problem can be solved efficiently by alternating optimization. Given current affinity matrix $\mathbf{C}$, we can optimize $\mathbf{A}$ via the Sinkhorn scaling algorithm, and then optimize $\mathbf{C}$ via stochastic gradient descent based on fixed $\mathbf{A}$.

We parameterize the affinity matrix $\mathbf{C}$ by a neural network, which takes paired sentence embeddings as its input and output their discrepancies according to their legal characteristics and semantics jointly. As illustrated in Figure 3, we model $\mathbf{C}$ as the combination of a rationale characteristic matrix $\mathbf{C}^r$ and a semantic matrix $\mathbf{C}^s$:

$$\mathbf{C} = \epsilon\mathbf{C}^r + \mathbf{C}^s, \tag{6}$$

where ϵ is a negative hyper-parameter to encourage the alignment of those sentence pairs which have the same rationale label by significantly reducing the transport between them. The $\mathbf{C}^s$ and $\mathbf{C}^r$ are constructed by the following steps.

3.3.1 Construction of the Semantic Matrix $\mathbf{C}^s$. IOT-Match constructs the semantic matrix $\mathbf{C}^s \in \mathbb{R}^{M\times N}$ to indicate the semantic distance between a cross-case sentence pair, i.e., $\mathbf{C}^s = \mathrm{dis}(\bar{\mathbf{s}}^X, \bar{\mathbf{s}}^Y)$, where function 'dis' measures the semantic distance of two sentence embeddings (e.g., Euclidean distance), $\bar{\mathbf{s}}_X$ and $\bar{\mathbf{s}}_Y$ respectively represent the contextual sentence embedding of legal case X and Y, which is obtained from a trainable two-layer MLP (projection layer in Figure 3) on the frozen sentence embeddings X and Y.

3.3.2 Construction of the Rationale Characteristic Matrix $\mathbf{C}^r$. The rationale characteristic matrix $\mathbf{C}^r$ indicates the rationales having the same legal characteristics, and the legal characteristics is categorized according to the rationale labels shown in Eq. (1). Taking two legal cases X and Y as the inputs, our IOT-Match predicts the rationale labels of their sentences, denoted as $\hat{\mathbf{r}}^X = \{\hat{r}_{x_m}\}_{m=1}^M$ and $\hat{\mathbf{r}}^Y = \{\hat{r}_{y_n}\}_{n=1}^N$, respectively, which is achieved by solving a sentence-level multi-class classification problem. Formally, given a legal case X, our IOT-Match would identify the legal characteristics of each sentence embedding x_m in X by calculating a probabilistic distribution over the four classes shown in Eq. (1):

$$\hat{r}_{x_m} = \arg\max_{k\in\{0,\cdots 3\}} P(r = k|x_m), \tag{7}$$

where $\{P(r = k|x_m)\}_{k=0}^3$ represent the distribution of the rationale labels conditioned on the sentence embedding x_m. In this work, we parameterize the distribution as follows:

$$\{P(r = k|x_m)\}_{k=0}^3 = \mathrm{softmax}(\mathbf{W}\mathbf{s}_{x_m}^{(L)} + \mathbf{b}), \tag{8}$$

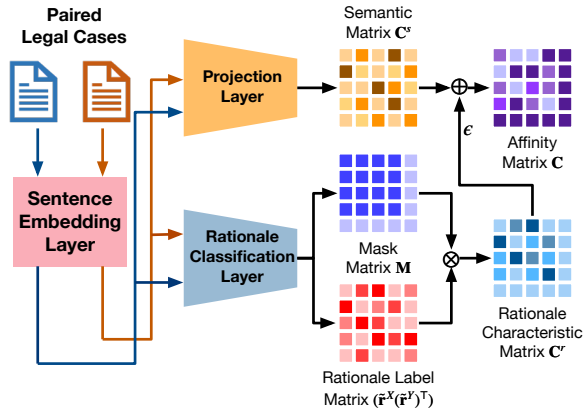


Figure 3: Illustration of the affinity matrix construction.

where the softmax converts a 4-dimensional vector to a distribution over four classes, matrix $\mathbf{W}$ and vector $\mathbf{b}$ are trainable parameters, and $\mathbf{s}_{x_m}^{(L)}$ is the output of a stacked of L-layer gated convolutional neural network [22] where the l-th layer is:

$$\mathbf{s}_{x_m}^{(l)} = \mathbf{s}_{x_m}^{(l-1)} + \text{conv}_1(\mathbf{s}_{x_m}^{(l-1)}) \otimes \sigma(\text{conv}_2(\mathbf{s}_{x_m}^{(l-1)})),$$

for $l = 1, \cdots, L$, and $\otimes$ denotes element-wise multiplication, conv_1 and conv_2 denote two dilate convolutional neural Network [49] with the same convolution kernel size. Note that the use of a stacked gated convolutional neural network enables the model to capture farther distances without increasing model parameters, which effectively addresses the issue caused by a large number of sentences in a legal case. $\sigma(\cdot)$ denotes a sigmoid gating function controlling which inputs $\text{conv}_1(\mathbf{s}_{x_m}^{(l-1)})$ of the current context are relevant. In the first layer, $\mathbf{s}_{x_m}^{(0)}$ is obtained by adding a trainable one-layer multi-layer perceptron on the frozen sentence embedding x_m.

Similarly, given a legal case Y, the legal characteristics of each sentence embedding y_n in Y can also be identified by classifying Y with the same sentence representations model and neural networks defined above. As a result, the rationale characteristic matrix $\mathbf{C}^r = [c^r_{mn}] \in \{0,1\}^{M\times N}$ can be defined to explicitly indicate whether two sentences have the same predicted legal characteristics:

$$\mathbf{C}^r = \mathbf{M} \otimes (\tilde{\mathbf{r}}^X(\tilde{\mathbf{r}}^Y)^\top), \tag{9}$$

where $\mathbf{M} \in \{0,1\}^{M\times N}$ is a mask matrix filtering out the padding sentences, $\tilde{\mathbf{r}}^X \in \{0,1\}^{M\times 4}$ and $\tilde{\mathbf{r}}^Y \in \{0,1\}^{N\times 4}$ are the rationale label matrix, whose rows are one-hot representations of $\hat{\mathbf{r}}^X$ and $\hat{\mathbf{r}}^X$. To incorporate Eq. (7) into Eq. (9) in a differentiable manner, we apply the Straight-Through Gumbel Trick [6] to derive $\tilde{\mathbf{r}}^X$ and $\tilde{\mathbf{r}}^Y$. Accordingly, $c^r_{mn} = 1$ means that the m-th sentence in X and the n-th sentence in Y are identified as rationales (i.e., $\hat{r}_{x_m} \neq 0$ and $\hat{r}_{y_n} \neq 0$) and they belong to the same rationale type ($\hat{r}_{x_m} = \hat{r}_{y_n}$).

3.4 Generating Candidate Explanations

As aforementioned, the optimal transport $\mathbf{A}^*$ indicates pro and con rationale pairs, and thus, can help to generate explanations (i.e., e's) to support matching results (i.e, z's). Following the work in [26], our IOT-Match exploits the existing pre-trained language model[2] to build three label-specific explanation generators, that is: $f_2 = \{G_z\}, z = 2, 1, 0$ respectively corresponds to matched, partially matched, and mismatched decisions as shown in Eq. (2).

The three generators are fine-tuned separately. For example, for $z = 0$, the data for fine-tuning G_0 is selected from the training corpus: $\mathcal{D}_0 = \{(X, Y, \hat{\mathbf{r}}^X, \hat{\mathbf{r}}^Y, e, z = 0)\} \subseteq \mathcal{D}$. Given each instance in $\mathcal{D}_0$, it is converted to the input sequence "$[x_{input}; y_{input}]$" which is a concatenation of two text sequences: $x_{input} = [T_1; x_1; \cdots; T_m; x_m]$ and $y_{input} = [T_1; y_1; \cdots; T_n; y_n]$, where x_i (y_j) is the sentence of the i-th (j-th) rationale in X (Y), T_i (T_j) is the special token indicating the rationale type [3], and m (n) is number of identified rationales. To fine-tune the parameters in G_0, the language modeling loss [37] that compares difference between the generated explanation $G_0([x_{input}; y_{input}])$ and the human-annotated explanation e is optimized. Similarly, G_1 and G_2 are fine-tuned based on corresponding subsets $\mathcal{D}_1$ and $\mathcal{D}_2$.

At explanation generation phase, given a tuple $(X, Y, \hat{\mathbf{r}}^X, \hat{\mathbf{r}}^Y)$, we feed the constructed input text sequence $[x_{input}; y_{input}]$ to G_0, G_1, and G_2, generating three candidate explanations $\hat{e}_0, \hat{e}_1$ and $\hat{e}_2$.

3.5 Matching Prediction

Instead of considering all of the sentences in the paired legal cases, IOT-Match learns the f_3 to conducts matching only based on the extracted rationales as well as the generated candidate explanations. This strategy makes the extracted rationales and generated explanations faithful to the matching predictions and avoids the negative impact of noise sentences on the matching results. Formally, given a paired legal case (X, Y), our IOT-Match would identify their relation by calculating a probabilistic distribution over the three classes shown in Eq. (2):

$$\{P(z = k|(X, Y)\}_{k=0}^{2} = \text{softmax}(\mathbf{W}[\hat{z}_0; \hat{z}_1; \hat{z}_2] + \mathbf{b}), \tag{10}$$

where $[;]$ concatenates vectors, and $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters. As for $\hat{\mathbf{z}}_i$ ($i \in \{0, 1, 2\}$), following the practice in [26, 54], the matching scores are computed based on the extracted rationales and the corresponding candidate explanations:

$$\hat{z}_i = \text{MLP}([\mathbf{s}_{r^X}; \mathbf{s}_{r^Y}; \mathbf{s}_{\hat{e}_i}]),$$

where $\mathbf{s}_{r^X}, \mathbf{s}_{r^Y}, \mathbf{s}_{\hat{e}_i}$ respectively denote the embeddings of rationales of X, Y, and the candidate explanation $\hat{e}_i$ which are obtained by tuning the BERT model[4]; MLP denotes a two-layer perceptron with sigmoid activation functions. Accordingly, our IOT-Match makes the final matching decision for a paired legal case (X, Y) as

$$\hat{z} = \arg\max_{k\in\{0,1,2\}} P(z = k|(X, Y)), \tag{11}$$

and outputs the explanation corresponding to the highest matching score at the same time.

3.6 Model Training

IOT-Match has parameters to determine during the training, including those in the pro and con rationales extraction (f_1), those in the candidate explanations generation (f_2), and those in the matching

[2]We adopt Chinese T5-PEGASUS model: https://github.com/ZhuiyiTechnology/t5-pegasus. Note that other pre-trained language models are also applicable.
[3]Six special tokens "[AI]", "[AO]", "[YI]", "[YO]","[ZI]", and "[ZO]" are defined, where A, Y, and Z stand for the key circumstance, constitutive elements of crime, and focus of disputes; I and O stand for pro and con. Non-rationale sentences were discarded.
[4]https://github.com/thunlp/OpenCLaP

Algorithm 1 Training Process of IOT-Match.

Require: Training set $\mathcal{D} = \{(X_i, Y_i, \mathbf{r}_i^X, \mathbf{r}_i^Y, \hat{\mathbf{A}}_i, z_i, e_i)\}_{i=1}^N$; mini-batch sizes n_1, n_2, n_3; trade-off coefficients $\epsilon, \gamma_1, \gamma_2, \gamma_3$; entropic regularizer coefficient γ; learning rates η_1, η_2, η_3.

▷ IOT-based Rationale Extraction
repeat
  Sample a mini-batch $\{(X_i, Y_i, \mathbf{r}_i^X, \mathbf{r}_i^Y, \hat{\mathbf{A}}_i)\}_{i=1}^{n_1}$ from $\mathcal{D}$
  Predict rationale labels $\hat{\mathbf{r}}_i^X, \hat{\mathbf{r}}_i^Y$ for (X_i, Y_i) {Eq. (8)}
  Calculate $\mathcal{L}_{\mathcal{R}}$ {Eq. (13)}
  Construct $\mathbf{C}^s, \mathbf{C}^r, \mathbf{C}$ and optimize $\mathbf{A}^*$ by Sinkhorn scaling
  Calculate $\mathcal{L}_{\mathcal{A}}$, with $\mathbf{A}^* = \mathbf{A}^*(\mathbf{C})$ {Eq. (14)}
  $\mathcal{L}_{f_1} = \sum_1^{n_1} \mathcal{L}_{\mathcal{R}} + \gamma_1 \mathcal{L}_{\mathcal{A}}$ {Eq. (12)}
  $\theta_{f_1} \leftarrow \theta_{f_1} - \eta_1 \bigtriangledown_{\theta_{f_1}} \mathcal{L}_{f_1}$
until convergence
return $\theta_{f_1}, \{\hat{\mathbf{r}}^X, \hat{\mathbf{r}}^Y\}$
▷ Generating Candidate Explanations
repeat
  Sample a mini-batch $\{(X_i, Y_i, \hat{\mathbf{r}}_i^X, \hat{\mathbf{r}}_i^Y, e_i)\}_{i=1}^{n_2}$ from $\mathcal{D}$
  Calculate $\mathcal{L}_{f_2}$ {Eq. (15)}
  Fine-tune three label-specific pre-trained language models.
until convergence
return $\theta_{f_2}, \{\hat{e}_0, \hat{e}_1, \hat{e}_2\}$
▷ Matching Prediction
repeat
  Sample a mini-batch $\{(X_i, Y_i, \hat{\mathbf{r}}_i^X, \hat{\mathbf{r}}_i^Y, z_i, e_i, \hat{e}_i)\}_{i=1}^{n_3}$ from $\mathcal{D}$
  Predict matching label $\hat{z}$ using predicted rationales candidate explanations $\hat{e}_0, \hat{e}_1, \hat{e}_2$ {Eq. (10)}
  Calculate $\mathcal{L}_{\mathcal{M}}$ {Eq. (17)}, $\mathcal{L}_{\mathcal{E}}$ {Eq. (18)}, and $\mathcal{L}_C$ {Eq. (19)}.
  $\mathcal{L}_{f_3} = \sum_1^{n_3} \mathcal{L}_{\mathcal{M}} + \gamma_3(\mathcal{L}_{\mathcal{E}} + \mathcal{L}_C)$ {Eq. (16)}
  $\theta_{f_3} \leftarrow \theta_{f_3} - \eta_3 \bigtriangledown_{\theta_{f_3}} \mathcal{L}_{f_3}$
until convergence
return θ_{f_3}

(f_3). These models parameters respectively denoted as $\theta_{f_1}, \theta_{f_2}, \theta_{f_3}$ are trained sequentially, and the output of the f_1 is used as the input of the f_2, and the output of the f_1 and f_2 are used as the input of the f_3. The training process of IOT-Match is illustrated in Algorithm 1. Specifically, in the f_1, the learning objective is defined to measure the loss of the pro and con rationales extraction:

$$\mathcal{L}_{f_1} = \sum\nolimits_{(X,Y,\mathbf{r}^X,\mathbf{r}^Y,\hat{\mathbf{A}})\in\mathcal{D}} \mathcal{L}_{\mathcal{R}} + \gamma_1 \mathcal{L}_{\mathcal{A}}, \tag{12}$$

where, for each legal case pair in the dataset, the loss function consists of two parts: the rationale identification loss $\mathcal{L}_{\mathcal{R}}$ and the affinity matrix loss $\mathcal{L}_{\mathcal{A}}$. The $\gamma_1 > 0$ is a hyper-parameter controlling their weights. The rationale identification loss $\mathcal{L}_{\mathcal{R}}$ is defined as the cross-entropy loss between the ground-truth rationale labels of each sentence and the corresponding predictions:

$$\begin{aligned}\mathcal{L}_{\mathcal{R}} = -\sum_{k=0}^{3}\Big(&\sum_{m=1}^{M}\delta(r_{x_m},k)\log(P(\hat{r}_{x_m}=k|x_m)) \\ &+\sum_{n=1}^{N}\delta(r_{y_n},k)\log(P(\hat{r}_{y_n}=k|y_n))\Big),\end{aligned} \tag{13}$$

where $\delta(r,k) = 1$ if $r = k$ else 0. The loss $\mathcal{L}_{\mathcal{A}}$ is based on the IOT problem in Eq. (5):

$$\mathcal{L}_{\mathcal{A}} = \mathrm{KL}(\hat{\mathbf{A}}||\mathbf{A}^*(\mathbf{C})) + \gamma_2 \sum_{m=1}^{M}\sum_{n=1}^{N}\delta(\hat{r}_{x_m},\hat{r}_{y_n})c_{mn}, \tag{14}$$

where the first term corresponds to the IOT problem that optimize the affinity matrix and the associated optimal transport to fit a small number of alignment labels (i.e., $\hat{\mathbf{A}}$). The second term is an unsupervised loss based on the predicted rationale labels, which explicitly regularizes the affinity matrix $\mathbf{C}$ to minimize the discrepancy between identical rationales and maximize the that between different rationales. Here, $\delta(\hat{r}_{x_m}, \hat{r}_{y_N}) = 1$ if $\hat{r}_{x_m} = \hat{r}_{y_N} \neq 0$ else 0, γ_2 is a coefficient to balance the supervised loss and the unsupervised loss.

In the f_2, its learning objective $\mathcal{L}_{f_2}$ is identical to that used in the fine-tuning phase of the pre-trained language models [26, 37, 53]:

$$\mathcal{L}_{f_2} = -\sum\nolimits_{(X,Y,\hat{\mathbf{r}}^X,\hat{\mathbf{r}}^Y,e)\in\mathcal{D}}\sum_{l=1}^{L}\log(P(s_l|\mathbf{s}_{1:l-1})), \tag{15}$$

where $\mathbf{s}$ stands for a sample sequence which contains L tokens, s_l denotes for the l-th token of $\mathbf{s}$, and $\mathbf{s}_{1:l-1}$ denotes the prefix of s_l.

In the f_3, the loss function consists of three parts:

$$\mathcal{L}_{f_3} = \sum\nolimits_{(X,Y,\hat{\mathbf{r}}^X,\hat{\mathbf{r}}^Y,z,e,\hat{e})\in\mathcal{D}} \mathcal{L}_{\mathcal{M}} + \gamma_3(\mathcal{L}_{\mathcal{E}} + \mathcal{L}_C), \tag{16}$$

where $\gamma_3 > 0$ is a coefficient to balance $\mathcal{L}_{\mathcal{M}}$, $\mathcal{L}_{\mathcal{E}}$ and $\mathcal{L}_C$. $\mathcal{L}_C$ is the cross-entropy loss between the ground-truth matching label z and the matching score of rationales and candidate explanations:

$$\mathcal{L}_{\mathcal{M}} = -\sum\nolimits_{k=0}^{2}\delta(z,k)\log\left(P(\hat{z}_k = k|(X,Y))\right), \tag{17}$$

where $\delta(z,k) = 1$ if $z = k$ else 0.

We also design two auxiliary tasks for learning a better representation for rationales and explanations. To ensure that the human-annotated explanation e accurately reflects the matching relation between rationales, the similarity between $[\mathbf{s}_{r^X}, \mathbf{s}_{r^Y}]$ and $\mathbf{s}_e$ should be larger than that between $[\mathbf{s}_{r^X}, \mathbf{s}_{r^Y}]$ and the generated explanation $\mathbf{s}_{\hat{e}_k}$. Therefore, the first task is designed as:

$$\begin{aligned}\mathcal{L}_{\mathcal{E}} = \sum_{k=0}^{2}\max\Big(0, &\cos(\mathrm{MLP}[\mathbf{s}_{r^X};\mathbf{s}_{r^Y}], \mathbf{s}_{\hat{e}_k}) \\ &-\cos(\mathrm{MLP}[\mathbf{s}_{r^X};\mathbf{s}_{r^Y}], \mathbf{s}_e)\Big),\end{aligned} \tag{18}$$

where MLP denotes a one-layer multi-layer perceptron. Moreover, inspired by the success of contrastive learning [14, 21, 29] and the observations in [26] that explanations with the same label tend to have the same form, and the form of explanations may be the noise for matching, the second auxiliary task is designed to avoid the classifier only using the form of explanations to infer the matching prediction. Specifically, the candidate explanations in current data are regarded as positive samples $\mathbf{s}_{\hat{e}_k}$, and the explanations with the same label in the mini-batch are regarded as negative samples $\mathbf{s}_{\hat{e}_k^l-}$. Then, the cosine similarity between rationales and positive/negative explanations are calculated and compared:

$$\begin{aligned}\mathcal{L}_C = \sum_{k=0}^{2}\sum_{l}\max\Big(0, &\cos(\mathrm{MLP}[\mathbf{s}_{r^X};\mathbf{s}_{r^Y}], \mathbf{s}_{\hat{e}_k^l-}) \\ &-\cos(\mathrm{MLP}[\mathbf{s}_{r^X};\mathbf{s}_{r^Y}], \mathbf{s}_{\hat{e}_k})\Big),\end{aligned} \tag{19}$$

where l is the number of negative samples.

4 ELAM AND ECAIL: NEW DATASETS FOR EXPLAINABLE LEGAL CASE MATCHING

To verify the effectiveness of our IOT-Match method, we construct a new Explainable Legal cAse Matching (ELAM) dataset that provides not only the ground-truth matching label for each legal case pair,

Table 1: Statistics of ELAM and eCAIL. Types of sentences include rationales and the others. #Rationale denotes the average number of rationales for each type per case; Prop. of pro and con denotes the average proportion of pro and con rationales per case.

	ELAM	eCAIL
Train/ Valid/ Test	4000/ 500/ 500	6000/ 750/ 750
Types of sentences	4	2
#Rationale	4.80/6.00/ 1.30	12.00
Prop. of pro and con	2.61:7.39	3.52:6.48
#Sentence per case	16.29	124.10
Length of explanation	176.34	165.39

but also manually-labeled rationales, their alignments, and natural language-based explanations of the matching decision.

To construct the ELAM dataset, we collect 8955 legal cases on "the obstruction of social management order crime" from Faxin[5]. Each case is already associated with several tags that provide some basic information such as application of the law, court name, judge year, trial or retrial, etc. During the pre-processing, we randomly sample 1250 cases as queries and construct a candidate pool for each query case. The cases in the candidate pool are retrieved according to their numbers of overlapped tags compared to the corresponding query case. To ensure the usability of the dataset, we remove those query cases (and their candidate pools) when the candidate pool retrieved less than 10 cases. In total, the final dataset contains 5000 legal case pairs. Finally, the basic information (e.g., the tags) is removed and some identity information is replaced with placeholders for privacy protection.

During the human annotation, each legal expert is provided a set of randomly selected legal case pairs. For each pair, a legal expert was asked to first annotate the rationale label for each sentence. Following the practices in [31], the rationale labels are 4-level: the key circumstances, the key constitutive elements of crime, the focus of disputes, and not a rationale. Then, the pro and con rationales (the alignment of rationales) were marked. Finally, the 3-level matching label was annotated. In addition, the legal experts were required to give a free-form natural language explanation for their matching decision based on the annotated rationales.

Besides ELAM, we also extended the CAIL (Challenge of AI in Law) 2021 dataset to adapt to the explainable legal case matching task. This dataset is created for the NLP competition in the law domain.[6] Each legal case in the original CAIL corpus is associated with several tags about the issue of private lending. In our extended CAIL (eCAIL), these tagged sentences are regarded as rationales, and others are regarded as unrelated sentences. The pro and con rationales correspond to the tagged sentences with identical labels and those with different labels, respectively. The same pre-processing as that of ELAM is conducted to construct the candidate legal case pairs. Because the tags in eCAIL data are the constitutive elements of crime for rationale sentences, we automatically assign a matching label for a pair of cases according to their tag-overlapping (overlapping > 10 means matching, overlapping < 1 means mismatching, else means partially matching). Similarly, as for the natural language-based explanation of the matching label, we concatenate all of the tags in a paired legal case. Some basic statistics of ELAM and eCAIL are listed in Table 1.

[5]https://www.faxin.cn

[6]We use the Fact Prediction Track data available at: http://cail.cipsc.org.cn/

5 EXPERIMENTS

In this section, we conduct experiments to answer the following research questions: **RQ1:** Can IOT-Match outperform state-of-the-art methods on legal case matching and text matching with explanations in terms of matching accuracy? **RQ2:** How good are the explanations produced by IOT-Match, including the extracted rationales and the generated natural language? **RQ3:** Can IOT-Match efficiently make use of limited rationale alignment labels?

The source code, ELAM and eCAIL datasets, and all experiments have been shared at: https://github.com/ruc-wjyu/IOT-Match.

5.1 Experimental Settings

5.1.1 Baselines and Evaluation Metrics. To the best of our knowledge, there exist few models that are exactly designed for explainable legal case matching. In the experiments, two types of text matching models are selected as baselines. The first type includes state-of-the-art legal case matching models without explanations:
1) **Sentence-BERT** [38] uses BERT pre-trained on the legal case corpus[7] to encode two cases and uses a MLP to conduct matching.
2) **Lawformer** [44] leverages a Longformer-based [4] pre-trained language model for Chinese legal long documents understanding.
3) **BERT-PLI** [41] uses BERT to capture paragraph-level semantic relations and then aggregates them with RNN and attention.
4) **Thematic Similarity** [8] segments two legal cases into paragraphs and computes the paragraph-level similarities. Maximum or average similarities are used for the overall matching prediction.

The second type of baselines includes the following matching models designed for short text matching with explanations:
1) **NILE** [26] adopts GPT2 to generate label-specific explanations for paired sentences, which has three variants that leverage different information to output matching score: **NILE (Ind)** only uses the generated explanation; **NILE (App)** uses the concatenation of input paired sentences and the generated explanation; and **NILE (Agg)** compares all the generated label-specific explanations.
2) **LIREx** [54] uses an attention mechanism to generate rationale-enabled explanations, which also involves selected explanations to conduct the sentence matching.

Note that both ELAM and eCAIL are in Chinese and do not have precedent information, we do not choose the precedent citation network-based methods [7, 8] as the baselines. Additionally, since NILE and LIREx can generate natural language explanations for matching, we compared IOT-Match with them in terms of explanation generation using identical pre-trained language models.

To evaluate the performance of rationale extraction, we also compare IOT-Match with the following state-of-the-art rationale extraction models designed for paired documents:
1) **MT-H-LSTM** [15] uses two bi-LSTMs to obtain sentence embeddings and predict the aligned sentences from document pairs.
2) **MLMC** [16] formulates the associative sentence extraction for

[7]https://github.com/thunlp/OpenCLaP

Table 2: Experimental results on ELAM and eCAIL test sets. † indicates the statistically significant difference between the performance of all baseline models and that of IOT-Match (p-value < 0.05).

Model types	Models	ELAM Acc. (%)	ELAM P. (%)	ELAM R. (%)	ELAM F1 (%)	eCAIL Acc. (%)	eCAIL P. (%)	eCAIL R. (%)	eCAIL F1 (%)
Legal case matching	Sentence-BERT [38]	68.83	69.83	66.88	67.20	71.33	70.83	71.21	70.98
	Lawformer [44]	69.91	72.26	68.34	69.18	70.67	70.20	70.55	69.91
	BERT-PLI [41]	71.21	71.22	71.23	70.88	70.66	70.05	70.54	70.18
	Thematic Similarity (avg) [8]	70.99	71.28	68.97	69.12	71.47	70.88	71.34	71.00
	Thematic Similarity (max) [8]	71.86	71.50	70.07	70.26	68.53	67.25	68.38	67.57
Short text matching with explanations	NILE (Agg) [26]	65.87	65.22	64.89	65.05	71.60	71.44	71.02	70.91
	NILE (App) [26]	68.90	68.90	66.87	67.32	72.53	71.97	71.93	71.95
	NILE (Ind) [26]	69.76	68.30	68.82	68.46	73.33	73.43	72.84	73.05
	LIREx [54]	68.18	68.22	67.34	67.66	70.53	69.68	70.40	69.94
Ours	IOT-Match	**73.87**†	**73.02**†	**72.41**†	**72.55**†	**82.00**†	**82.10**†	**81.92**†	**81.90**†

paired documents as a problem of table filling, in which a matrix is constructed to show whether the sentences are related or not.
3) **DecAtt** [35] adopts attention to indicate the alignments between cross-case sentences. To make fair comparisons, the sentence encoder in DecAtt is set to be identical to that of in IOT-Match.

Different metrics are adopted to evaluate the different modules of IOT-Match. As for rationales extraction and matching prediction, Accuracy, Precision, Recall, and F1 are used. As for natural language explanation generation, the ROUGE score is used because the task is formulated as the Seq2Seq text generation.

5.1.2 Hyper-parameter settings. All of the hyper-parameters in IOT-Match are tuned using grid search on the validation set with Adam [24]. In the rationale extraction, the learning rate η_1 is tuned between $\{1e{-}4, 1e{-}3\}$; the batch size n_1 is tuned among $\{32, 64, 128\}$; γ_1 is tuned between $[1, 10]$ and γ_2 is tuned between $[0.1, 1.0]$; the alignment threshold τ for ELAM and eCAIL are tuned between $[1e-3, 1e-2]$ and $[1e-3, 5e-3]$, respectively; and the entropic regularizer γ is tuned among $[0.1, 1.0]$; the affinity matrix coefficient ϵ is tuned among $\{0, -10, -50, -100, -200\}$. In the natural explanation generation, the hyper-parameters are set according to those reported in [42]: the learning rate η_2 is set as $2e-5$; the batch size n_2 is set as 2; In the matching, the learning rate η_3 is tuned between $\{2e-5, 2e-4\}$; the batch size n_3 is tuned between $\{4, 8\}$, and γ_3 is tuned among $\{1, 10, 20\}$.

5.2 Matching Accuracy (RQ1)

We first study the matching performance of our proposed IOT-Match. Table 2 presents the matching performances of IOT-Match and the baselines in terms of four evaluation metrics on ELAM and eCAIL. All the methods are trained ten times and the averaged results are reported. Based on the results, we summarize our observations as follows: (1) IOT-Match consistently and significantly outperforms all of the baselines on two datasets in terms of all metrics, indicating the effectiveness of IOT-Match in enhancing the matching accuracy. Note that the legal cases in eCAIL are extremely lengthy (on average over 100 sentences per legal case). IOT-Match achieves over 10% improvements in terms of all four metrics on eCAIL, further verifying its effectiveness in the matching of long-form legal cases. (2) Compared to short text matching with explanation models which involve all sentences in a paired legal case during the matching, IOT-Match enjoys the advantages from the extracted rationales and achieves consistent improvements on two datasets. The result indicates that the rationale extraction module in IOT-Match accurately identified the rationales and filtered out the noise sentences from legal cases. (3) Compared to existing legal case matching models that cannot provide matching explanations, IOT-Match also achieves consistent improvements on both datasets. The results indicate that the natural language explanations generated by IOT-Match are helpful for legal case matching.

Table 3: Plausibility of extracted rationales on ELAM and eCAIL test sets in terms of extraction accuracy.

Models	ELAM Acc. (%)	eCAIL Acc. (%)
MT-H-LSTM [15]	68.91	95.18
MLMC [16]	68.37	95.30
DecAtt [35]	83.09	94.33
OT [36]	83.09	90.97
IOT-Match	**86.82**	**96.26**

5.3 Quality of Rationales and Explanations (RQ2)

The major superiority of IOT-Match compared to existing legal case matching models is that IOT-Match is able to extract rationales and generate explanations for the matching prediction. In this subsection, we conduct experiments to assess the quality of the extracted rationales and the generated natural language explanation by IOT-Match. Following [18], we adopt plausibility and faithfulness as the metrics. Plausibility measures how well the explanation aligns with human annotations, and faithfulness measures the degree to which the explanation influences the corresponding predictions.

5.3.1 Quality of the extracted rationales. In terms of **plausibility**, we compare the rationales extracted by IOT-Match and baseline models with human annotations on ELAM and eCAIL. As shown in Table 3, the rationales extracted by IOT-Match are more consistent with human annotations, especially on the ELAM dataset

Table 4: Faithfulness of the extracted rationales and the generated explanation on ELAM and eCAIL test sets. The column "Input" denotes IOT-Match with different inputs.

	ELAM				CAIL			
Input	**Acc. (%)**	**P. (%)**	**R. (%)**	**F1 (%)**	**Acc. (%)**	**P. (%)**	**R. (%)**	**F1 (%)**
$a\backslash r$	65.01	64.24	63.29	63.35	72.40	71.78	72.30	71.87
a	68.83	69.83	66.88	67.20	71.33	70.83	71.21	70.98
r	70.35	70.06	68.94	69.22	72.67	72.29	72.53	72.08
e	71.27	70.47	70.71	70.58	68.47	68.52	65.67	66.05
$a\backslash r + e$	69.98	70.75	68.79	69.51	79.20	79.63	69.45	79.14
$a + e$	73.65	**73.29**	**73.29**	**73.26**	71.33	70.83	71.21	70.98
$r + e$	**73.87**	73.02	72.41	72.55	**82.00**	**82.10**	**81.92**	**81.90**

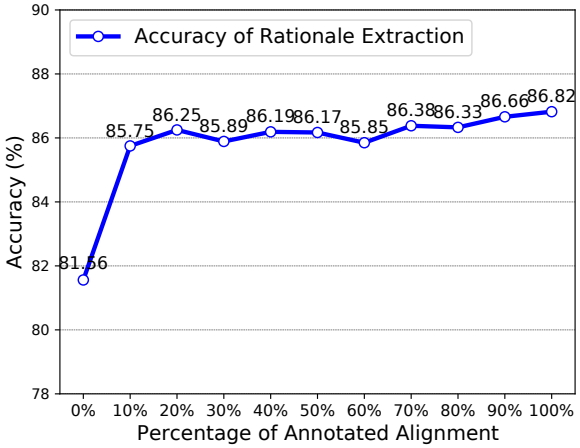


Figure 4: Rationale extraction accuracy of IOT-Match w.r.t. different percentages of labeled alignments.

where the rationales are more diverse (three types of rationales). According to the results illustrated in the figure, IOT-Match outperforms MLMC [16], MT-H-LSTM [15] and DecAtt [35] by about 24.0%, 25.0%, and 4.0%, respectively on ELAM. We also compare the original IOT-Match with a modified one with IOT ablated but forward OT kept, denoted as OT in Table 3. Table 3 shows that the extraction accuracy drops if we remove IOT from IOT-Match. The results indicate the effectiveness of IOT in learning the adaptive cross-case sentence affinity and predicting the rationale alignment.

In terms of **faithfulness**, we conduct experiments to measure the degree to which the extracted rationales influence the final matching. Specifically, we test the matching performance of IOT-Match with explanations and IOT-Match without explanations respectively under three conditions: using all sentences as the input (respectively denoted as "IOT-Match ($a + e$)" and "IOT-Match (a)"), using rationale extracted by IOT-Match as the input (respectively denoted as "IOT-Match ($r + e$)" and "IOT-Match (r)"), and using sentences except those extracted by IOT-Match as the input (respectively denoted as "IOT-Match ($a\backslash r + e$)" and "IOT-Match ($a\backslash r$)"). From the results reported in Table 4, we find that the rationales extracted by IOT-Match play a critical role in legal case matching. Specifically, if the extracted rationales are removed from a model's input (IOT-Match($a\backslash r + e$) or IOT-Match($a\backslash r$)), the matching accuracy of the model drops dramatically. In addition, in eCAIL where the legal cases are extremely lengthy, if all sentences are used as a model's input (IOT-Match($a + e$) or IOT-Match(a)), the model's accuracy still drops to some extent because of the noise from other sentences. On ELAM, the performance of using rationales as the only input is competitive with that of using all sentences. This result indicates the rationales extracted by IOT-Match already provide sufficient legal semantics for case matching. Based on the above analysis, we conclude that IOT-Match is capable of accurately extracting faithful rationales for legal case matching.

5.3.2 Quality of the Generated Explanation. In terms of **plausibility**, we compare the natural language explanation generated by IOT-Match to those generated by NILE [26] and LIREx [54]. Since both ELAM and eCAIL have human-annotated explanations for the matching labels, the popular metrics in machine translation such as ROUGE-1, ROUGE-2, and ROUGE-L are used to evaluate the plausibility. As shown in Table 5, the natural language explanations generated by IOT-Match are more consistent with human annotations than those generated by NILE and LIREx, especially on the eCAIL where IOT-Match outperformed NILE [26] and LIREx [54] at least by 7.9% and 2.5% across all metrics, respectively. Moreover, we also conduct human evaluations to test the quality of the generated explanations. Following [54], we randomly sampled 50 examples respectively from ELAM and eCAIL, and ask two annotators to answer the questions that whether the generated explanation and the label explanation convey the same meaning. Each annotator was provided with the context (legal cases, rationales, explanations), and asked to label them as 1 if they agree to the question, or 0 otherwise. As shown in Table 6, IOT-Match obtains a high relevance score between the generated explanations and label explanations. The result verifies the effectiveness of IOT-Match in generating plausible explanations.

In terms of **faithfulness**, we conduct experiments to measure the degree to which the generated explanation influences the final matching. Specifically, we compare the performance among IOT-Match using the rationales only (IOT-Match(r)), using the explanation only (IOT-Match(e)), and using rationales and explanations (IOT-Match($r + e$)). From the results reported in Table 4, we find: (1) on both ELAM and eCAIL, IOT-Match ($r + e$) performs the best, indicating that the natural language explanations generated by IOT-Match contributed to the matching prediction; (2) IOT-Match (e) performs better than IOT-Match (r) on ELAM, verifying the faithfulness of the generated explanation. The result also indicates that the explanations on ELAM are more sufficient for the matching prediction than the extracted rationales. (3) IOT-Match (e) performs worse than IOT-Match (r) on eCAIL. We analyze the reasons and find that the labeled explanations on eCAIL are the concatenations of the rationale sentences. Such labeled explanations are not coherent enough and may harm the generated explanations.

5.4 Robustness under limited labels (RQ3)

One advantage of IOT-Match is its capability of learning to extract the pro and con rationales from human-labeled rationale alignments

Table 5: Plausibility of generated explanations on ELAM and eCAIL test sets in terms of ROUGE scores.

Models	ELAM ROUGE-1 (%) $z=2$ $z=1$ $z=0$	ELAM ROUGE-2 (%) $z=2$ $z=1$ $z=0$	ELAM ROUGE- L (%) $z=2$ $z=1$ $z=0$	eCAIL ROUGE-1 (%) $z=2$ $z=1$ $z=0$	eCAIL ROUGE-2 (%) $z=2$ $z=1$ $z=0$	eCAIL ROUGE- L (%) $z=2$ $z=1$ $z=0$
NILE [26]	73.40 70.96 69.93	58.47 55.57 56.08	69.87 65.70 66.84	73.40 70.96 69.93	58.47 55.57 56.08	69.87 65.70 66.84
LIREx [54]	74.15 71.61 70.97	59.78 56.36 56.88	70.89 66.22 67.41	80.35 74.36 74.46	72.00 67.35 63.58	76.84 74.28 66.98
IOT-Match	**75.55 73.64 75.18**	**60.97 58.25 61.84**	**72.79 68.85 72.54**	**83.54 76.59 91.21**	**76.48 69.46 87.03**	**80.37 76.16 86.91**

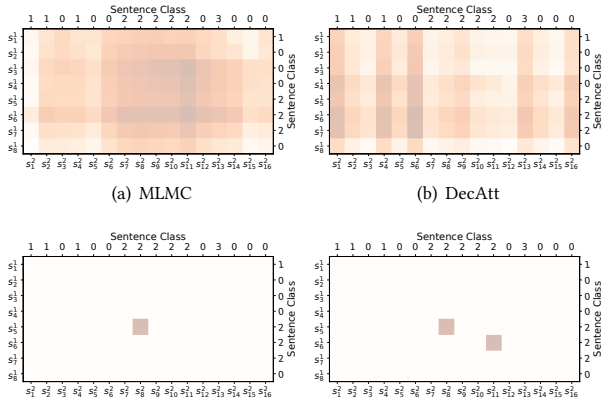
Table 6: Human evaluations of the explanation quality over 50 randomly sampled data from ELAM and eCAIL by two annotators with the inter-rater agreement of 0.95.

	NILE [26]	LIREx [54]	IOT-Match
ELAM	35	41	**46**
eCAIL	36	38	**44**

in a semi-supervised manner, because, in real practice, manually labeling the rationale alignments is expensive and time-consuming.

We conduct experiments to test the rationale extraction accuracy w.r.t. different amounts of labeled alignments. Specifically, we configure IOT-Match to extract rationales given different ratios of labeled alignments $\hat{\mathbf{A}}$ in Eq. (5) (from 0% to 100% where 0% means no labels available, and 100% means fully supervised learning). Figure 4 illustrates the extraction accuracy w.r.t. the ratio of labeled alignments on ELAM data. We find that IOT-Match shows competitive performances when only 10%~20% of the labeled alignments are involved in learning. The results indicate that with only a small fraction of the alignment labels, IOT-Match can still learn the cross-case sentence-level affinity matrix $\mathbf{C}$ with high accuracy, and accurately extract the pro and con rationales.

Figure 5 shows the predicted rationale alignments for an example legal case pair (used in Figure 2) from ELAM test set. The models are trained when only 10% of the alignment labels are used. From the results, we find that MLMC and DecAtt output dense alignments (Figure 5(a) and (b)), which are not accurate (ground-truth alignments are shown in Figure 5(d)) and are hard to be explained. In contrast, IOT-Match not only accurately identifies and aligned the rationales (Figure 5(c)), but also is explainable due to its sparseness. The results verify that IOT-Match is able to robustly and accurately extract and align the rationales when the alignment labels are insufficient.



(a) MLMC (b) DecAtt

(c) IOT-Match (d) Human labeled alignments

x_8: The defendant, Ruan **, in collaboration with others, forged a residential identity card, which constitutes the crime of forging identity documents and should be punished according to law.
y_5: The court held that the defendant Zhou ** gang forged residential identity cards, disturbing public order, his behavior has constituted the crime of forging identity documents.

x_{11}: The defendant Ruan ** voluntarily and truthfully confessed to his crime, admitted the alleged criminal acts, was willing to accept the punishment, and could be leniently punished according to the law.
y_6: The public prosecutor's charge was found guilty, the defendant Zhou ** confessed truthfully after his return to the case, willing to accept the punishment, according to the law to be mitigated; prepayment of fines, the discretion to mitigate the punishment.

(e) Aligned sentence pairs in the example cases (translated from Chinese)

Figure 5: Rationale alignments of an example case pair from ELAM test set where 10% of alignments are used for training. (a), (b), and (c): predicted rationale alignments; (d): human labeled alignments; (e): two labeled sentence pairs.

6 CONCLUSION

This paper proposes a novel inverse optimal transport-based model called IOT-Match for explainable legal case matching. IOT-Match is not only able to extract the pro and con rationales and generates natural language explanations for legal case matching, but is also robust to alignment label insufficiency. A new dataset is created to facilitate the study of explainable legal case matching. Comprehensive experimental results showed that IOT-Match consistently outperformed state-of-the-art baselines in terms of matching accuracy. The empirical analysis verified that the extracted rationales and the generated explanations are not only consistent with human annotations but also faithful to the final matching prediction.

ACKNOWLEDGMENTS

This work was funded by the National Key R&D Program of China (2019YFE0198200), the National Natural Science Foundation of China (61872338, 61832017, 62106271, 62102420), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, the Mainland-Hong Kong Joint Funding Scheme (MHP/001/19) from the Innovation and Technology Commission (ITC) of Hong Kong, Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, and Public Policy and Decision-making Research Lab of Renmin University of China.

REFERENCES

[1] David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-Wasserstein Alignment of Word Embedding Spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1881–1890.

[2] David Alvarez-Melis, Stefanie Jegelka, and Tommi S. Jaakkola. 2019. Towards Optimal Transport with Global Invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan (Proceedings of Machine Learning Research, Vol. 89)*. PMLR, 1870–1879. http://proceedings.mlr.press/v89/alvarez-melis19a.html

[3] Katie Atkinson, Trevor J. M. Bench-Capon, and Danushka Bollegala. 2020. Explanation in AI and law: Past, present and future. *Artif. Intell.* 289 (2020), 103387. https://doi.org/10.1016/j.artint.2020.103387

[4] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv preprint* abs/2004.05150 (2020). https://arxiv.org/abs/2004.05150

[5] Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G Conrad, Enrico Francesconi, et al. 2012. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law* 20, 3 (2012), 215–319.

[6] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv preprint* abs/1308.3432 (2013). https://arxiv.org/abs/1308.3432

[7] Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2020. *Hier-SPCNet: A Legal Statute Hierarchy-Based Heterogeneous Network for Computing Legal Case Document Similarity.* Association for Computing Machinery, New York, NY, USA, 1657–1660.

[8] Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2020. Methods for Computing Legal Document Similarity: A Comparative Study. *CoRR* abs/2004.12307 (2020). arXiv:2004.12307 https://arxiv.org/abs/2004.12307

[9] Adrien Bibal, Michael Lognoul, Alexandre De Streel, and Benoît Frénay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* 29, 2 (2021), 149–169.

[10] Saptarshi Chakraborty, Debolina Paul, and Swagatam Das. 2020. Hierarchical clustering with optimal transport. *Statistics & Probability Letters* 163 (2020), 108781.

[11] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: "Preparing the Muppets for Court'". In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*. Association for Computational Linguistics, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[12] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph Optimal Transport for Cross-Domain Alignment. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1542–1553. http://proceedings.mlr.press/v119/chen20e.html

[13] Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Improving Sequence-to-Sequence Learning via Optimal Transport. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=S1xtAjR5tX

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607. http://proceedings.mlr.press/v119/chen20j.html

[15] 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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7000–7011. https://aclanthology.org/2020.emnlp-main.569

[16] Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. 2021. Argument Pair Extraction via Attention-guided Multi-Layer Multi-Cross Encoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6341–6353. https://aclanthology.org/2021.acl-long.496

[17] Marco Cuturi. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf

[18] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4443–4458. https://aclanthology.org/2020.acl-main.408

[19] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *CoRR* abs/1711.01134 (2017). arXiv:1711.01134 http://arxiv.org/abs/1711.01134

[20] Arnaud Dupuy, Alfred Galichon, and Yifei Sun. 2016. Estimating matching affinity matrix under low-rank constraints. *ArXiv preprint* abs/1612.09585 (2016). https://arxiv.org/abs/1612.09585

[21] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[22] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 1243–1252. http://proceedings.mlr.press/v70/gehring17a.html

[23] Xin Jiang, Hai Ye, Zhunchen Luo, WenHan Chao, and Wenjia Ma. 2018. Interpretable Rationale Augmented Charge Prediction System. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Santa Fe, New Mexico, 146–151. https://aclanthology.org/C18-2032

[24] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

[25] Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, and Aditya Singh. 2011. Similarity analysis of legal judgments. In *Proceedings of the 4th Bangalore Annual Compute Conference, Compute 2011, Bangalore, India, March 25-26, 2011*. ACM, 17. https://doi.org/10.1145/1980422.1980439

[26] Sawan Kumar and Partha Talukdar. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8730–8742. https://aclanthology.org/2020.acl-main.771

[27] Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. 2019. Learning to Match via Inverse Optimal Transport. *J. Mach. Learn. Res.* 20 (2019), 80:1–80:37. http://jmlr.org/papers/v20/18-700.html

[28] Liting Liu, Wenzheng Zhang, Jie Liu, Wenxuan Shi, and Yalou Huang. 2021. Interpretable Charge Prediction for Legal Cases based on Interdependent Legal Information. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*. IEEE, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533902

[29] Yixin Liu and Pengfei Liu. 2021. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 1065–1072. https://aclanthology.org/2021.acl-short.135

[30] Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining Relationships Between Scientific Documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2130–2144. https://aclanthology.org/2021.acl-long.166

[31] 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. *Information Retrieval (IR)* 2 (2021), 22.

[32] Akshay Minocha, Navjyoti Singh, and Arjit Srivastava. 2015. Finding Relevant Indian Judgments Using Dispersion of Citation Network. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15 Companion)*. Association for Computing Machinery, New York, NY, USA, 1085–1088. https://doi.org/10.1145/2740908.2744717

[33] Alfredo López Monroy, Hiram Calvo, Alexander F. Gelbukh, and Georgina García Pacheco. 2013. Link Analysis for Representing and Retrieving Legal Information. In *Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 7817)*. Springer, 380–393. https://doi.org/10.1007/978-3-642-37256-8_32

[34] Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1938–1952. https://aclanthology.org/2020.emnlp-main.153

[35] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2249–2255. https://aclanthology.org/D16-1244

[36] Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* 11, 5-6 (2019), 355–607.

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[38] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://aclanthology.org/D19-1410

[39] Manavalan Saravanan, Balaraman Ravindran, and Shivani Raman. 2009. Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law* 17, 2 (2009), 101–124.

[40] Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. 2021. Learning from the Best: Rationalizing Predictions by Adversarial Information Calibration. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 13771–13779. https://ojs.aaai.org/index.php/AAAI/article/view/17623

[41] 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 the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org, 3501–3507. https://doi.org/10.24963/ijcai.2020/484

[42] Jianlin Su. 2021. *T5 PEGASUS - ZhuiyiAI*. Technical Report.

[43] Cédric Villani. 2009. *Optimal transport: old and new*. Vol. 338. Springer.

[44] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents. *CoRR* abs/2105.03887 (2021). arXiv:2105.03887 https://arxiv.org/abs/2105.03887

[45] Hongteng Xu, Dixin Luo, and Lawrence Carin. 2019. Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/6e62a992c676f611616097dbea8ea030-Paper.pdf

[46] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. 2019. Gromov-Wasserstein Learning for Graph Matching and Node Embedding. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 6932–6941. http://proceedings.mlr.press/v97/xu19b.html

[47] Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. 2018. Distilled Wasserstein learning for word embedding and topic modeling. *Advances in Neural Information Processing Systems* 31 (2018).

[48] 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 the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1854–1864. https://aclanthology.org/N18-1168

[49] Fisher Yu and Vladlen Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. http://arxiv.org/abs/1511.07122

[50] Weijie Yu, Chen Xu, Jun Xu, Liang Pang, Xiaopeng Gao, Xiaozhao Wang, and Ji-Rong Wen. 2020. Wasserstein Distance Regularized Sequence Representation for Text Matching in Asymmetrical Domains. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2985–2994. https://aclanthology.org/2020.emnlp-main.239

[51] Weijie Yu, Chen Xu, Jun Xu, Liang Pang, and Ji-Rong Wen. 2022. Distribution Distance Regularized Sequence Representation for Text Matching in Asymmetrical Domains. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 721–733.

[52] Yiming Zeng, Ruili Wang, John Zeleznikow, and Elizabeth A. Kemp. 2005. Knowledge Representation for the Intelligent Legal Case Retrieval. In *Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, Melbourne, Australia, September 14-16, 2005, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 3681)*. Springer, 339–345. https://doi.org/10.1007/11552413_49

[53] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 11328–11339. http://proceedings.mlr.press/v119/zhang20ae.html

[54] Xinyan Zhao and V. G. Vinod Vydiswaran. 2021. LIREx: Augmenting Language Inference with Relevant Explanations. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 14532–14539. https://ojs.aaai.org/index.php/AAAI/article/view/17708

[55] Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 1250–1257. https://aaai.org/ojs/index.php/AAAI/article/view/5479