Mind Your Neighbours: Leveraging Analogous Instances for Rhetorical Role Labeling for Legal Documents

Santosh T.Y.S.S, Hassan Sarwat, Ahmed Abdou, Matthias Grabmair
School of Computation, Information, and Technology;
Technical University of Munich, Germany
{santosh.tokala, hassan.sarwat, ahmed.abdou, matthias.grabmair}@tum.de

Abstract

Rhetorical Role Labeling (RRL) of legal judgments is essential for various tasks, such as case summarization, semantic search and argument mining. However, it presents challenges such as inferring sentence roles from context, interrelated roles, limited annotated data, and label imbalance. This study introduces novel techniques to enhance RRL performance by leveraging knowledge from semantically similar instances (neighbours). We explore inference-based and training-based approaches, achieving remarkable improvements in challenging macro-F1 scores. For inference-based methods, we explore interpolation techniques that bolster label predictions without re-training. While in training-based methods, we integrate prototypical learning with our novel discourse-aware contrastive method that work directly on embedding spaces. Additionally, we assess the cross-domain applicability of our methods, demonstrating their effectiveness in transferring knowledge across diverse legal domains.

Keywords: Rhetorical Role Labeling, Prototypical Learning, Contrastive Learning, Interpolation

1. Introduction

In an era of rapid digitalization and exponential growth of legal case volumes, the demand for automated systems to assist legal professionals in tasks like extracting key case elements, summarizing cases, and retrieving relevant cases has surged (Zhong et al., 2020). At the core of these tasks lies Rhetorical Role Labeling (RRL), which involves assigning functional roles to the sentences in the document such as preamble, factual content, evidence, reasoning, etc. Legal documents, characterized by their extensive length, lengthy sentences with unusual word order, frequent cross-references, extensive citation usage, and intricate lexicon, often feature uncommon expressions from everyday language and borrowed terms from various languages to the extent that they are referred to as a sub-language of legalese (Chalkidis et al., 2022; Haigh, 2013).

The task of RRL faces several distinctive challenges. Firstly, contextual dependencies, influenced by surrounding sentences and the case's context, are pivotal in discerning rhetorical role of each sentence, distinguishing RRL as a *sequential sentence classification* task. Secondly, the intertwining nature of rhetorical roles further complicates the task. For instance, the rationale behind a judgment (Ratio of the decision) often overlaps with Precedents and Statutes, necessitating a nuanced understanding of these roles' intricate distinctions (Bhattacharya et al., 2021). Thirdly, obtaining extensive annotated data for specialized domains like law is expensive, requiring expert annotators. Lastly, certain rhetorical roles are disproportionately represented in the dataset, leading to significant class imbalance (Malik et al., 2022; Bhattacharya et al., 2021). Traditional up/down sampling methods struggle to fully address this challenge due to the task's nature, which involves sequences of sentences at the document level.

Initially RRL task is formulated as sentence classification, treating each sentence in isolation (Ahmad et al., 2020; Walker et al., 2019). Researchers later adopted it as sequential sentence classification, addressing contextual dependencies between sentences (Bhattacharya et al., 2021; Ghosh and Wyner, 2019; Malik et al., 2022; Kalamkar et al., 2022). They introduced a two-level hierarchical model, encoding sentences independently at the lower level and contextualizing them with neighbouring sentences at the higher-level. While this approach effectively addressed the first challenge of RRL, other challenges remain unaddressed. Recently, Santosh et al. 2023 aimed to address data scarcity through data augmentation, but methods like word deletion, sentence swapping and back-translation could introduce noise and disrupt coherence. However, this approach did not effectively address label imbalance and intricate role intertwining.

In this work, we hypothesize that harnessing knowledge from semantically and contextually similar instances can provide valuable insights to grasp a broader context and reveal underlying rare patterns. This can enhance the understanding of complex label-semantics relationships, improve nuanced label assignments and equip the model to handle less common labels, thus addressing the distinctive challenges of RRL. We ex-

LREC-COLING 2024, pages 11296–11306
20-25 May, 2024.

plore two approaches for harnessing this knowledge: one directly at inference time without additional parameters or re-training (Sec. 4), and the other during training by incorporating auxiliary loss constraints (Sec. 5). In the inference-based approaches, we interpolate the label distribution predicted by a model with the distribution derived from analogous instances in the training dataset, employing nearest neighbor-based, single, and multiple prototype-based methodologies. These methods enhance performance, particularly on more challenging macro-F1 scores, without requiring retraining. For training-based approaches, we integrate contrastive and prototypical learning which operate directly on the embedding space, leveraging neighborhood relationships. Additionally, we introduce a novel discourse-aware contrastive loss to address the contextual nature of the task. Our experimental results on four datasets from the Indian Jurisdiction validate our proposed methods.

While it is common to develop models for specific courts or domains due to unique vocabulary, complex linguistic structures and specific writing styles, such specialization can hinder the adaptability of these models beyond their original context. In rhetorical role labeling, models might memorize context-specific vocabulary rather than understanding the underlying semantics, making cross-domain applications challenging (Savelka et al., 2021). In such cases, developing a model for a new context typically requires annotating a new dataset, which can be expensive. In our work, we assess the cross-domain generalizability of our methods and observe that they enhance model's ability to transfer across different legal domains compared to a baseline model lacking these auxiliary techniques (Sec. 6).

2. Related Work

Rhetorical Role Labeling Initial efforts of RRL aimed to facilitate summarization tasks (Saravanan et al., 2008; Farzindar and Lapalme, 2004). Saravanan et al. 2008 employed Conditional Random Fields on hand-crafted features, to identify seven rhetorical roles in Indian state High Court documents. Savelka and Ashley 2018 categorized text into functional segments (Introduction, Background, Analysis, and Footnotes) and issue-specific segments (Analysis and Conclusion) using CRF on a corpus of US trade secret and cybercrime decisions. Walker et al. 2019 adopted feature-based methods for segmenting U.S. Board of Veterans' Appeals decisions. Nejadgholi et al. 2017 focused on identifying factual and non-factual sentences in Canadian immigration case documents, using FastText embeddings for query-oriented search engine application.

Recently, deep learning-based classification have been applied to this task in various contexts, such as Japanese documents (Yamada et al., 2019), Indian Supreme Court documents (Bhattacharya et al., 2021; Ghosh and Wyner, 2019; Malik et al., 2022; Kalamkar et al., 2022). These methods adopt hierarchical approaches to account for the sequential sentence classification nature of the task, drawing context from surrounding sentences. This has been the defacto architecture for this task, with modifications ranging from word embeddings initially (Bhattacharya et al., 2021; Ghosh and Wyner, 2019) to BERT based contextualized embeddings recently (Malik et al., 2022; Kalamkar et al., 2022). Recently, Santosh et al. 2023 reformulated the task as span-level sequential classification that segment the document into sets of contiguous sequence of sentences (spans) and assign them labels. In our work, we make use of the Indian Supreme Court corpus from prior research, proposing algorithms to effectively enhance their performance leveraging the knowledge from analogous neighbourly instances both at inference time without re-training, and also during training. Recently, Savelka et al. 2021 investigated the transferability of rhetorical segmentation models across seven jurisdictions and six languages, including Canada, the US, Czech Republic, Italy, Germany, Poland and France. In this study we also examine cross-domain performance on Indian Supreme Court documents across different legal contexts.

Leveraging Neighborhood Information Utilizing neighborhood information offers two pathways: one during inference and the other during training. *Inference-time methods*, commonly applied in few-shot classification, facilitate label assignment based on proximity to training examples without any re-training (Snell et al., 2017; Yang and Katiyar, 2020). Various techniques include employing all examples to identify nearest neighbors for the final label assignment (Yang and Katiyar, 2020) and constructing prototypes based on examples of the same label (Snell et al., 2017), among others. This concept has gained widespread attention as retrieval-augmented models in various tasks, including language modeling (Khandelwal et al., 2019; Zheng et al., 2021), machine translation (Zheng et al., 2021)named entity recognition (Wang et al., 2022b) and multi-label text classification (Wang et al., 2022a).

On the training side, methods like contrastive learning have been applied in self-supervised representation learning (Gao et al., 2021), wherein neighbour constraints are enforced in the embedding space through data augmentation. More recently, they have been extended to supervised learning scenarios using instances with the same label as neighbors (Khosla et al., 2020). Another approach is prototypical learning (Ding et al., 2020), which designates representative prototypes for each class as guiding points to enforce neighborhood constraints on data instances. In this study, we harness both training and inference-based neighbour learning strategies. Additionally, we explore the their capabilities in cross-domain scenarios, within the context of the RRL task.

3. Task, Datasets, Baseline

Task Given a judgment document $x = \{x_1, x_2, \ldots, x_m\}$ with m sentences as the input, where $x_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ represents the i^{th} sentence containing n tokens and x_{jp} refers to the p^{th} token in the j^{th} sentence, the task of rhetorical role labeling is to predict sequence of $l = \{l_1, l_2, \ldots, l_m\}$ where l_i is the rhetorical role corresponding to sentence x_i and $l_i \in$ L which is set of predefined rhetorical role labels.

Data We experiment on four datasets - **(i) Build** (Kalamkar et al., 2022) comprises judgments from Indian supreme court, high court, and district courts. It includes publicly available train and validation splits, with 184 and 30 documents respectively with a total of 31865 sentences (an average of 115 per document). These documents pertain to tax and criminal law cases and are annotated with 13 rhetorical role labels, including 'None'. Given the absence of a public test dataset, we utilize the training dataset for both training and validation, evaluating performance on the validation partition. **(ii) Paheli** (Bhattacharya et al., 2021) features 50 judgments from the Supreme Court of India across five domains: Criminal, Land and Property, Constitutional, Labour and Industrial, and Intellectual Property Rights, annotated with 7 rhetorical roles. They have total of 9380 sentences with an average of 188 per document. **(iii) M-CL / (iv) M-IT** (Malik et al., 2022) encompasses judgments from the Supreme Court of India, High Courts, and Tribunal courts. It includes two subsets: M-CL, comprising 50 documents related to Competition Law, and M-IT, with 50 documents related to Income Tax cases. Both subsets are annotated with 7 rhetorical role labels. M-CL has 13,328 sentences (avg. of 266 per document) and M-IT has a total of 7856 sentences (avg. of 157 per document). We split (at document level) Paheli/M-CL/M-IT into 80% train, 10% validation, and 10% test set.

Baseline All of our experiments in this study are built on top of the Hierarchical Sequential Labeling Network, which served as a baseline in prior works (Kalamkar et al., 2022; Santosh et al., 2023). Initially, each sentence x_i is encoded independently using a BERT model (Kenton and Toutanova, 2019) to derive token-level representations $z_i = \{z_{i1}, z_{i2}, \ldots, z_{in}\}$. These representations are passed through a Bi-LSTM layer (Hochreiter and Schmidhuber, 1997), followed by an attention pooling layer (Yang et al., 2016), to yield sentence representations $s = \{s_1, s_2, \ldots, s_m\}$.

$$u_{it} = \tanh(W_w z_{it} + b_w) \tag{1}$$

$$\alpha_{it} = \frac{\exp(u_{it} u_w)}{\sum_s \exp(u_{is} u_w)} \quad \& \quad s_i = \sum_{t=1}^{n} \alpha_{it} u_{it} \tag{2}$$

Here, W_w, b_w, and u_w are trainable parameters. The sentence representations s are passed through Bi-LSTM layer to obtain contextualized representations $c = \{c_1, c_2, \ldots, c_m\}$ that encode contextual information from surrounding sentences. Finally, the contextual representations c are passed through a Conditional Random Field layer that predicts the best sequence of labels.

4. RQ 1: Leveraging the Neighbourhood at Inference

In this section, we leverage the knowledge from semantically similar training instances directly during inference without extra training overhead. We interpolate the label distribution predicted by the baseline model with the distribution derived from the training instances similar to the test instance. This overcomes the problem of memorizing/learning rare patterns implicitly in the model parameters, thus enhancing the model's ability to handle long-tail cases (classes with few instances or rare patterns in frequent classes) especially in limited data settings. We explore three different methods to obtain the distribution from similar training instances.

4.1. Methods

4.1.1. Interpolation with kNN

In this method, we construct a datastore of training instances and then retrieve the nearest neighbours to the test instance for computing the interpolated label distribution during the inference.

Datastore Construction After training, we obtain contextualized representation c_i of every sentence in each document of the training set using the trained model. We construct the datastore by a single forward pass over each training document. The datastore $\{K, V\}$ is the set of all contextualized representation-rhetorical label pairs constructed from all the training examples D as:

$$\{K, V\} = \{(c_i, l_i) | \forall x_i \in x, \forall l_i \in l, (x, l) \in D\} \quad (3)$$

Interpolation During inference time, we query the datastore using the contextualized representation of every sentence in the test document, to find the k-nearest neighbours N according to the euclidean distance. Then, we derive the distribution of labels p_{kNN} using labels of the retrieved neighbours based on softmax of their negative distances, while aggregating probability mass for each label across all its occurrences in the retrieved neighbours (labels that do not appear in the retrieved N have zero probability). Intuitively, the closer a neighbor is to the test instance, the larger its weight is.

$$p_{kNN}(l_i|x, x_i) \propto \sum_{(k,v) \in N} \mathbb{1}_{l_i=v} \exp(\frac{-d(c_i, k)}{\tau}) \quad (4)$$

τ denotes the temperature hyperparameter and d(.) indicates euclidean distance. Finally, we interpolate the $p_{baseline}(l_i|x, x_i)$ with $p_{kNN}(l_i|x, x_i)$ as:

$$\begin{aligned} p_{final}(l_i|x, x_i) = \lambda p_{baseline}(l_i|x, x_i) + \\ (1 - \lambda) p_{kNN}(l_i|x, x_i) \end{aligned} \quad (5)$$

where λ makes a balance between derived p_{kNN} and $p_{baseline}$ obtained from the trained model.

4.1.2. Interpolation with Single Prototype

Instead of storing all the training instances in the datastore, we seek to store one prototype for each label, which captures the essential semantics of various sentences for each rhetorical role, significantly reducing the datastore's memory footprint. To create a prototype for each label, we calculate the average of the contextualized embeddings of sentences that share the same rhetorical role. Intuitively these prototypes can be assumed to be the center of clusters for different labels, surrounded by sentences expressing the same label in the embedding space. The interpolation process closely resembles the kNN approach (Eq. 5), with the key difference being that interpolation directly involves the prototypes, rather than a prior retrieval step.

4.1.3. Interpolation with Multiple Prototypes

Instead of using a single prototype for each rhetorical role, we suggest the use of multiple prototypes for each label. This choice is driven by the fact that instances with the same rhetorical role can exhibit distinct variations in expression, resulting in diverse contextual embeddings scattered across the embedding space. Averaging these embeddings into a single prototype might diminish specificity. Utilizing multiple prototypes allows us to effectively capture the intricate viewpoints within each label. To accomplish this, we cluster the instances belonging to each rhetorical role using k-means, yielding multiple prototypes for each label from the k centroids. The interpolation step remains similar (Eq. 5), involving all these multiple prototypes without any retrieval step.

4.2. Experiments

4.2.1. Implementation Details

We follow the hyperparameters for baseline as described in Kalamkar et al. 2022. We use the BERT base model to obtain the token encodings. We employ a dropout of 0.5, maximum sequence length of 128, LSTM dimension of 768, attention context dimension of 200. We sweep over learning rates {1e-5, 3e-5, 5e-5. 1e-4, 3e-4} for 40 epochs with Adam optimizer (Kingma and Ba, 2014) to derive the best model based on validation set performance. For all our inference variants, we carry a grid search over the interpolation factor (λ) in increments of 0.1 in the range of [0,1] to choose the best model based on Macro-F1 on validation set. For KNN and multiple prototypes, we vary k over powers of 2 from 8 till 256.

4.2.2. Results

In Table 1, we present the macro-F1 and micro-F1 scores for both the baseline and the interpolation variants. We observe a significant improvement when using kNN interpolation across all datasets, particularly in the more challenging macro-F1 metric, which accounts for label imbalances. On the other hand, single prototype interpolation mitigates memory footprint issue of kNN by storing one representation per rhetorical role but leads to performance degradation compared to kNN. This decline results from oversimplification, as a single prototype may struggle to capture the diverse aspects within each rhetorical role, particularly when instances of the same label are dispersed across the embedding space. This is evident in the Paheli dataset, where no improvement over the baseline is observed. Interpolation with multiple prototypes balances memory efficiency and label variation capture. While it slightly underperforms kNN interpolation in Paheli and M-CL datasets, it outperforms kNN in Build and M-IT. This can be attributed to a smoothing effect that reduces noise or human label variations in the kNN-based approach, particularly evident in datasets with low inter-annotator agreements (Build and M-IT). These results affirm our hypothesis that straightforward interpolation using training set examples during inference can boost the performance of rhetorical role classifiers.

	Build		Paheli		M-CL		M-IT	
	mac.F1	mic.F1	mac.F1	mic.F1	mac.F1	mic.F1	mac.F1	mic.F1
Baseline	60.20	79.13	62.43	66.02	59.51	67.04	70.76	70.50
+ KNN	62.92	81.04	66.53	70.82	63.14	73.02	72.16	71.62
+ Single Proto	61.23	80.12	62.43	66.02	61.42	71.64	71.97	71.08
+ Mutli Proto	63.23	81.96	65.36	70.02	62.73	72.78	72.82	72.46

Table 1: Performance of interpolation methods on four datasets. mac.F1: macro-F1, mic.F1: micro-F1

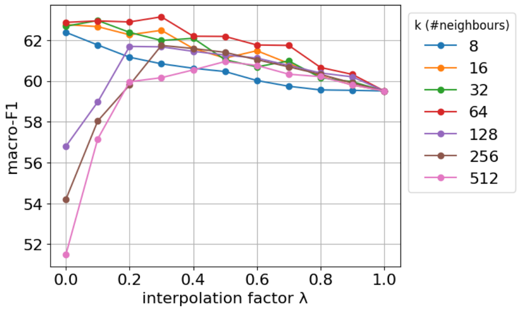


Figure 1: Sensitity to hyperparameters - kNN (M-CL) λ = 0: interpolation only, λ = 1: baseline only

Sensitivity of interpolation In Figure 1, we present the macro-F1 score for the M-CL dataset using kNN interpolation, while varying the interpolation coefficient λ and the number of neighbors 'k'. Here, λ values of 0 and 1 correspond to predictions solely from interpolation and the baseline model, respectively. We observe that performance initially improves as 'k' increases, signifying that incorporating more neighbors boosts confidence by including closely similar examples. However, performance starts to decline with higher 'k', which can be attributed to a large number of neighbours introducing noise with low inter-annotator agreement, suggesting a need for a addressing this task a multi-label classification. On the other hand, reducing λ consistently enhances performance, particularly for lower k, showcasing the model's capacity to rely solely on semantically similar instances for label prediction. With higher k, we notice a decline in performance at lower λ values beyond a certain optimal point, which is related to the label variation problem exacerbated by a larger number of neighbours. Similar trends are observed with other interpolations.

5. RQ 2: Leveraging the Neighbourhood at Training

We leverage the knowledge from neighbour instances directly at the training process to improve the performance. We explore three methods: contrastive learning, single prototypical learning, and multi prototypical learning. These techniques draw inspiration from the same principles as their inference-time counterparts but serve as auxiliary loss constraints during training. Their primary aim is to improve the discriminative capability of embeddings by highlighting differences between instances with distinct rhetorical roles and similarities among instances sharing the same label.

While the task-specific classification loss focuses on mapping contextualized representations to label outputs with supervision on individual instances, the methods in this section directly operate on embeddings in latent space. They exploit the interplay among instances to establish effective discriminative decision boundaries, serving as a form of regularization.

5.1. Methods

5.1.1. Contrastive Learning

Contrastive learning aims to bring an anchor point closer to related samples while pushing it away from unrelated samples in the embedding space. In a supervised setting, samples with the same/different labels are considered related/unrelated with respect to an anchor (Khosla et al., 2020). The loss is calculated as follows:

$$L^{cont=-\frac{1}{N^2}\sum_{i,j}} \frac{\exp(\delta(c_i,c_j)d(c_i,c_j))}{\sum_{j'}\exp(1-\delta(c_i,c_{j'}))d(c_i,c_{j'})} \tag{6}$$

$$d(c_i,c_j) = \frac{1}{(1+\exp(\frac{c_i}{|c_i|}\cdot\frac{c_j}{|c_j|}))} \tag{7}$$

where $\delta(c_i,c_j)$ denotes 1 if c_i and c_j have same rhetorical label, else 0, N denotes batch size.

Lengthy legal documents limits the number of documents that can be accommodated in a single batch and this raises concerns about having enough positive samples for the minority class instances within a batch for effective contrasting. To overcome this limitation, we utilize a **memory bank** (Wu et al., 2018), where we progressively reuse encoded representations from previous batches to compute the contrastive loss. In practice, we maintain a fixed-size representation queue for each rhetorical role. As new representations corresponding to specific labels are generated, they are enqueued into the respective queue with their gradients detached. If the queue size for a label exceeds the maximum limit, the oldest element is dequeued. When computing the contrastive loss, we use the same equation 7. However, in addition to the current batch instances, we employ all the representations stored in the memory bank for contrasting purposes, using them as both positives and negatives, based on the anchor point's label.

To incorporate the concept of context from surrounding sentences into contrastive learning, we introduce a novel **discourse-aware contrastive loss**. This is based on the idea that sentences in close proximity within a document, sharing the same label, should exhibit a stronger proximity in the embedding space compared to sentences with the same label but positioned farther apart in the document. To implement this concept, we introduce a penalty inversely proportional to the absolute difference in their positions. In particular, we impose a higher penalty on positive sentence pairs that are closer in the document, encouraging them to be closer in the embedding space than pairs originating from greater distances within the document. The discouse-aware loss is as follows:

$$L^{cont} = -\frac{1}{N^2}\sum_{i,j}\frac{\text{exp}(\beta(i,j)\delta(c_i,c_j)d(c_i,c_j))}{\sum_{j'}\text{exp}(1-\beta(i,j)\delta(c_i,c_{j'}))d(c_i,c_{j'})} \tag{8}$$

$$\beta(i,j) \propto \frac{1}{|j-i|} \tag{9}$$

where β represents a penalty that considers positional information. When c_i and c_j come from different documents, such as cross-document positives/negatives from the memory bank or across the batch, we apply the lowest possible penalty, considering c_i as the farthest sentence relative to in-document positives. We incorporate this additional contrastive loss alongside the classification loss during training.

5.1.2. Single Prototypical Learning

While contrastive learning guides instances to adjust their positions in the embedding space relative to other instances , prototypical learning employs specific prototypes for each label in the embedding space which act as specific guiding points (Ding et al., 2020). Specifically, we randomly initialize one prototype for each label as $z = \{z_1, z_2, \ldots, z_k\}$, where k denotes the number of rhetorical roles. These prototypes undergo fine-tuning during the training and we apply distance-related constraints from both the prototype's and the sample's perspectives to guide their relationships. (i) Prototype centric view (pcv): aims to bring samples S_j belonging to label j closer to the corresponding prototype z_j, while simultaneously pushing away samples of other labels, denoted as S'_j, from this prototype. (ii) Sample centric view (scv): In this view, the representation c_j with label j is drawn closer to its designated prototype z_j, while pushing away from other prototypes $Z'_j = z - z_j$. These two views are represented in loss as:

$$L_j^{pcv} = \text{-}\frac{1}{N}(\sum_{c_p \in S_j}\text{log}(d(z_j,c_p)) + \sum_{c_i \in S'_j}\text{log}(1\text{–}d(z_j,c_i))) \tag{10}$$

$$L_j^{scv} = \text{-}\frac{1}{K}(\text{log}(d(z_j,c_j)) + \sum_{z_p \in Z'_j}\text{log}(1\text{–}d(z_p,c_j))) \tag{11}$$

These both views shape the embedding space by aligning prototypes with their corresponding samples, forming distinct clusters of different labels, each centered around a specific prototype vector.

5.1.3. Multi Prototypical Learning

Instead of using a single prototype for each label, this approach employs multiple prototypes for each label to capture the diverse variations within the sentences of the same label. To implement this, a set of M prototypes per label is randomly initialized and a diversity loss (Zhang et al., 2022) is integrated to penalize prototypes of the same label if they are too similar to each other. This ensures that prototypes of the same label are distributed across the embedding space, capturing the multifaceted nuances under each label. The Sample Centric View is also modified to ensure that each sample is in close proximity to at least one prototype among all the prototypes of the same class.

$$L_k^{div} = \sum_{\substack{q \neq r \\ z_q, z_r \in Z_k}} \text{max}(0, z_q \cdot z_r - \theta) \tag{12}$$

$$\begin{aligned} L_j^{scv} = &- \min_{z_q \in Z_k} \text{log}(d(z_q,c_j) + \\ &\frac{1}{(k-1)M}\sum_{z_p \in Z'_k}\text{log}(1 - d(z_p,c_j)) \end{aligned} \tag{13}$$

where z_q, z_r are prototypes of same label k. Sample c_j belongs to label k. θ is the similarity threshold. Z_k and Z'_k represent the set of prototypes corresponding to the label k and those of labels other than k, respectively.

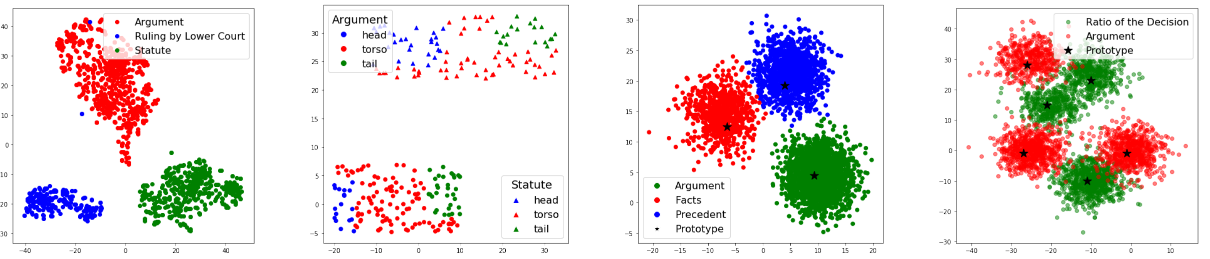
5.2. Experiments

5.2.1. Implementation Details

We use the same training setup as described in Sec. 4.2.1. We conduct grid-search for size of memory bank per label and number of prototypes in multi-prototypical learning in powers of 2 from [32,512] and [4,256] respectively using the validation set performance.

	Build		Paheli		M-CL		M-IT	
	mac.F1	mic.F1	mac.F1	mic.F1	mac.F1	mic.F1	mac.F1	mic.F1
Baseline	60.20	79.13	62.43	66.02	59.51	67.04	70.76	70.50
+ Contrastive	64.55	83.54	68.06	71.91	62.24	72.42	73.41	73.53
+ Contrastive + MB	66.51	83.29	71.76	72.69	63.14	72.72	72.22	72.46
+ Disc. Contr.	66.37	83.81	71.99	73.85	66.94	73.02	72.23	74.01
+ Disc. Contr. + MB	66.48	83.67	71.19	73.28	64.72	72.36	72.85	73.05
+ Single Proto.	66.01	81.45	69.94	71.09	64.42	71.52	72.59	71.98
+ Multi Proto.	66.35	83.05	71.38	72.92	65.91	73.57	73.02	74.13
+ Disc. Contr. **+ Single Proto.**	67.02	83.91	74.28	73.86	65.87	72.12	72.50	72.1
+ Disc. Contr. **+ Multi Proto.**	67.21	83.65	75.52	76.34	68.66	74.59	73.14	72.22

Table 2: Results on four datasets for methods leveraging neighbourhood during training (RQ2). Contr., Disc., MB, Proto. indicates Contrastive, Discouse-aware, Memory Bank and Prototypical respectively.



(a) Contrastive (b) Disc.-aware Contr. (c) Single-Prototypical (d) Multi-prototypical

Figure 2: t-SNE visualizations of different models on M-CL dataset. Disc.: Discourse, Contr.: Contrastive. head, torso and tail in Disc.-aware Contr. plot indicate the relative position of the sentence in a document.

5.2.2. Results

Table 2, shows that incorporating contrastive loss improves performance across all datasets. Furthermore, the discourse-aware contrastive loss, which leverages relative position to organize embeddings, enhances performance, supporting our hypothesis that sentences with the same label and in close proximity in the document should be closer in the embedding space. Augmenting the contrastive loss with a memory bank further enhances performance, particularly in macro-F1, benefiting sparse classes. However, the degree of improvement is less or negative in the discourse-aware variant. This can be due to the positional factor, as additional sentences from other documents retrieved from the memory bank are placed at the end of the document, leading to smaller penalization factors and contributing only marginally to the loss. Overall, the discourse-aware contrastive model emerges as the most effective among the contrastive variants.

The single prototypical variant performs comparably to the best contrastive variant and outperforms the baseline. This demonstrates that specific guiding points through prototypes can effectively aggregate knowledge from neighboring instances. Moreover, multiple prototypes further improve performance, highlighting the need to capture multifaceted nuances. These results suggest that the addition of respective losses can eliminate the need to design specific memory banks to expose the model to large batches for effective guidance from neighbors in contrastive learning.

Finally, combining the discourse-aware contrastive variant with both single and multiple prototype variants yields further improvement, highlighting the complementarity between these approaches. These results suggest that deriving supervisory signals from interactions among training instances can be an effective strategy for addressing the class imbalance problem, particularly in low-data settings.

Qualitative Analysis: To examine the impact of our auxiliary loss functions on the learned representations, we employ t-SNE (Hinton and Roweis, 2002) to project the high-dimensional latent space hidden states obtained by the model in Fig. 2. In the case of contrastive learning, we observe that sentences with the same label form distinct clusters. With the addition of discourse-aware contrastive loss, samples with the same label in a specific document adhere to the positional constraint, aligning with our hypothesis that samples sharing a label and closer in the discourse se-

Train ↓	Test →	Paheli	M-CL	M-IT
	Random	19.10	7.87	9.12
Paheli	Baseline	62.43	56.98	57.31
	Disc. Contr.	71.99	56.54	57.40
	Single Proto.	69.94	58.30	59.92
	Multi Proto.	71.38	57.47	59.48
	DC + Single Pr	74.28	62.27	60.33
	DC + Multi Pr	75.52	60.89	60.61
M-CL	Baseline	54.71	59.51	63.08
	Disc. Contr.	54.04	66.94	62.98
	Single Proto.	57.48	64.42	60.23
	Multi Proto.	56.10	65.91	61.62
	DC + Single Pr	59.95	65.87	63.92
	DC + Multi Pr	57.89	68.66	62.37
M-IT	Baseline	52.97	58.83	70.76
	Disc. Contr.	51.89	57.16	72.23
	Single Proto.	51.57	58.58	72.59
	Multi Proto.	52.85	58.70	73.02
	DC + Single Pr	51.03	57.23	72.50
	DC + Multi Pr	51.77	56.99	73.14

Table 3: Macro-F1 scores of our methods across three datasets. The column 'train' indicates the source dataset on which the model is trained and each of the dataset columns indicates the target test dataset. Scores in grey indicates the in-domain performance (trained and tested on same dataset). {DC, Disc. Contr.} : Discourse-aware contrastive, {Pr., Proto.} : Prototypical

quence should be positioned closer in the embedding space compared to those farther apart. In single prototypical learning, prototypes occupy the centers of corresponding sentences, forming distinctive manifolds. Similarly, multi-prototypical learning captures multifaceted aspects with prototypes dispersed across the embedding space, each prototype serving as the center for respective samples. These visualizations affirm the effectiveness of our learning methods.

6. RQ 3: Cross-Domain Generalizability

To evaluate how well our proposed methods can transfer across different domains, we train the model on one dataset (source) and assess its performance on other datasets (target) in a blind zero-shot manner. We use the Paheli, M-CL, and M-IT datasets, which span diverse domains but share a same 7 rhetorical label space. The resulting Macro-F1 scores are presented in Table 3.

Naturally, models trained and tested on the same domain outperform those trained on different domains (e.g., baseline model trained and tested on Paheli achieves a Macro-F1 of 62.43, whereas trained on M-CL and tested on Paheli achieves 54.71). Interestingly, the baseline model shows an ability to transfer knowledge from one domain to another, outperforming random[1] guessing across all datasets. While discourse-aware contrastive model improves in-domain performance, it marginally reduces cross-domain performance across all datasets compared to the baseline (e.g., Disc. Contr. trained on M-CL and tested on Paheli achieves a Macro-F1 of 54.04, while the baseline with the same setup achieves 54.71). This can be attributed to the model capturing domain-specific features while minimizing distances between similar instances in contrastive learning. In contrast, single and multi-prototypical models enhance cross-domain transfer compared to the baseline, except when trained on M-IT. This indicates prototypical learning acts as a more robust guiding point, preventing overfitting to noisy neighbors as in contrastive models. Between the two, single prototype tend to perform better, due to its single representation being agnostic to domain-specific variations and encapsulating core characteristics, making it more adept in cross-domain scenarios. Furthermore, coupling discourse-aware contrastive with prototypical models boosts cross-domain performance, except when trained on M-IT. This behaviour of M-IT may be attributed to marginal in-domain improvements, leading to overfitting on domain-specific features limiting cross-domain generalization. This prompts questions about selection of optimal source dataset for improved performance on target datasets, warranting further investigation. For instance, to test on Paheli with baseline, training on M-CL yields a Macro-F1 of 54.71, while on M-IT yields 52.97. Additionally, exploring joint training with multiple datasets could shed light on their impact on both in-domain source and unseen target datasets.

7. Conclusion

In this paper, we have demonstrated the potential for enhancing the performance of rhetorical role classifiers by leveraging knowledge from neighbours, semantically similar instances. Interpolation with kNN and multiple prototypes at the inference time have shown promising improvements, especially in addressing the challenging issue of label imbalance, without requiring re-training. Additionally, our approach of incorporating neighbourhood constraints during training with our proposed discourse-aware contrastive learning and prototypical learning has demonstrated improvements. Combining both methods has boosted it further, indicating their complementary nature. Notably, the prototypical methods have proven to

[1] Random choices are based on the training set's label distribution (uniform distribution lead to further lower scores). These are averaged over 10 runs.

be robust, showcasing performance gains even in cross-domain scenarios, generalizing beyond the domains they were trained on.

8. Limitations

One constraint in the current task formulation is that it restricts assigning a single label to each sentence, which may not fully account for the complexity of lengthy sentences that can encompass multiple rhetorical roles. To address this limitation, an alternative approach could involve reformulating the task as multi-label classification, enabling each sentence to be associated with more than one gold-standard rhetorical role. Another avenue for exploration is to shift from sentence-level segmentation towards a finer-grained approach at the phrase or sub-sentence level, necessitating the assignment of rhetorical roles to each phrase or sub-sentence while specifying the dependency relations between these segments (Tokala et al., 2023).

It's important to acknowledge that while our cross-domain experiments have provided valuable insights into model generalizability, these evaluations have primarily focused on datasets originating from Indian courts, covering various domains within this single jurisdiction. The observed improved performance across these datasets could potentially be attributed to shared country-specific vocabulary and legal conventions. To ensure the robustness and generalizability in a broader context, it is imperative to expand the assessment to encompass diverse legal contexts across different countries and regions, where legal documents from may exhibit significant linguistic and structural variations.

9. Ethics Statement

The scope of this work is to provide new methods along with corresponding experiments to drive research forward in rhetorical role labeling, which is a pivotal task constituting the inaugural step in the legal document processing pipeline. Our experiments have been carried out on four publicly available datasets from different Indian courts. Though these decisions are not anonymized and contain the real names of the involved parties, we do not foresee any harm incurred by our experiments. We believe that our research contributes positively to the broader goals of advancing legal NLP and the development of AI-driven tools for legal professionals. By enhancing the automation of rhetorical role labeling, we can streamline legal document analysis and significantly benefit the legal field.

10. Bibliographical References

Syed Rameel Ahmad, Deborah Harris, and Ibrahim Sahibzada. 2020. Understanding legal documents: classification of rhetorical role of sentences using deep learning and natural language processing. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 464–467. IEEE.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2021. Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, pages 1–38.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Ning Ding, Xiaobin Wang, Yao Fu, Guangwei Xu, Rui Wang, Pengjun Xie, Ying Shen, Fei Huang, Hai-Tao Zheng, and Rui Zhang. 2020. Prototypical representation learning for relation extraction. In *International Conference on Learning Representations*.

Atefeh Farzindar and Guy Lapalme. 2004. Letsum, an automatic text summarization system in law field. JURIX.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6894–6910. Association for Computational Linguistics (ACL).

Saptarshi Ghosh and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. *Legal Knowledge and Information Systems: JURIX*, page 3.

Rupert Haigh. 2013. *Legal english*. Routledge.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429.

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

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2022. Semantic segmentation of legal documents via rhetorical roles. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 153–171.

Isar Nejadgholi, Renaud Bougueng, and Samuel Witherspoon. 2017. A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In *JURIX*, pages 125–134.

TYS S Santosh, Philipp Bock, and Matthias Grabmair. 2023. Joint span segmentation and rhetorical role labeling with data augmentation for legal documents. In *European Conference on Information Retrieval*, pages 627–636. Springer.

M Saravanan, Balaraman Ravindran, and S Raman. 2008. Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Jaromir Savelka and Kevin D Ashley. 2018. Segmenting us court decisions into functional and issue specific parts. In *JURIX*, pages 111–120.

Jaromir Savelka, Hannes Westermann, Karim Benyekhlef, Charlotte S Alexander, Jayla C Grant, David Restrepo Amariles, Rajaa El Hamdani, Sébastien Meeùs, Aurore Troussel, Michał Araszkiewicz, et al. 2021. Lex rosetta: transfer of predictive models across languages, jurisdictions, and legal domains. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 129–138.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Yaswanth Sri Sai Santosh Tokala, Sai Saketh Aluru, Anoop Vallabhajosyula, Debarshi Kumar Sanyal, and Partha Pratim Das. 2023. Label informed hierarchical transformers for sequential sentence classification in scientific abstracts. *Expert Systems*, 40(6):e13238.

Vern R Walker, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning. *ASAIL@ ICAIL*, 2385.

Ran Wang, Xinyu Dai, et al. 2022a. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679.

Shuhe Wang, Xiaoya Li, Yuxian Meng, Tianwei Zhang, Rongbin Ouyang, Jiwei Li, and Guoyin Wang. 2022b. k nn-ner: Named entity recognition with nearest neighbor search. *arXiv preprint arXiv:2203.17103*.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2019. Neural network based rhetorical status classification for japanese judgment documents. In *Legal Knowledge and Information Systems*, pages 133–142. IOS Press.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. 2022. Protgnn: Towards self-explaining graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9127–9135.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive nearest neighbor machine translation. 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)*, pages 368–374.

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