Graph-based Logical Reasoning for Legal Judgement

Stanford CS224N Custom Project
Mentor: Tathagat Verma
(Optional) Sharing project: No

Ein Jun
Department of Computer Science
Stanford University
ejun26@stanford.edu

Abstract

Despite the fact that legal judgement prediction (LPJ) specifically entails the utilization of reasoning capacities, the papers which address the specific problem generally do not directly address the task of logical reasoning- and rather suffer from the lack of such mechanisms within the proposed models. Thus, this paper applies the method of modeling logical relations through a graph network to the task of LPJ, which not only supplements the lack of explanation that current model results suffer from, but also integrate logical understanding to models, stepping beyond the currently existing methods in the discourse.

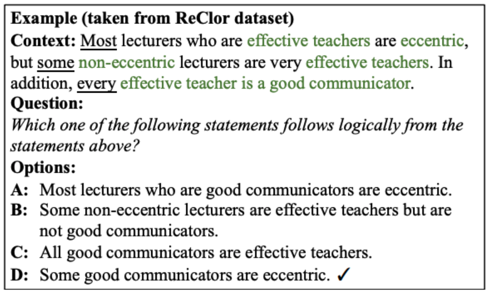
1 Background

Introduction and Related Work As in the recent years, datasets targeting machine reading comprehension (MRC) have been evolving to introduce more complex and intricate patterns of text. Particularly, state-of-the-art models such as BERT or RoBERTa, while capturing bias on simpler sets with high accuracy, reported performance approximately equivalent to random guessing on these more difficult and logically complex datasets ((Yu et al., 2020)). As a prime example that emulates this development, and a approach that motivates the approach of this paper, Chen et al. (2022) attempts to address the issue of how existing methods of logical reasoning MRC target either entity awareness or discourse based information, but overlook how the elements of the text interact through hierarchical relations which are key to identifying logical relations ((Yu et al., 2020) ; (Liu et al., 2020); (Wang et al., 2021) ; (Huang et al., 2021); (Ouyang et al., 2021)). Therefore, the paper proposes a holistic graph network (HGN) which targets the collective consideration hierarchical features within the text, and of the relations between different levels of granularity (the questions, paragraphs, sentences, and entities) within the text.

As previously mentioned, the papers which address the specific problem of legal judgement prediction generally do not directly address the task of logical reasoning. While after the publishing of Chalkidis et al. (2019), there have been more papers which introduced neural models to the discourse, as opposed to simply relying on linear models, these papers also have not addressed the issue of logical reasoning directly. Specifically, in Chalkidis et al. (2019), it was mentioned that while neural methods outperform feature-based models, they provide no justification for their predictions, and following studies employing neural methods have also been limited to elucidating which parts of the text affect predictions the most based upon the attention scores.

Reverting attention to the legal domain, the recent use of graph-based neural methods in LJP have been only employed to survey the relations between cases, modeling the cases or law articles as nodes of a graph, and predicting their relations through edges ((Huang et al., 2023), (Khatri et al., 2023)). Given the importance and contingency that patterns of hierarchical relations within the text hold for legal documents, applying this method of logical reasoning to the LJP task is a promising endeavor.

Approach Despite the ECHR dataset being the dataset for which a significantly larger portion of benchmark evaluations are available, the introduction of neural models to the domain of legal judgement prediction was a fairly recent phenomena, with Chalkidis et al. (2019) being considered



Example (taken from ReClor dataset)
Context: Most lecturers who are effective teachers are eccentric, but some non-eccentric lecturers are very effective teachers. In addition, every effective teacher is a good communicator.
Question:
Which one of the following statements follows logically from the statements above?
Options:

A: Most lecturers who are good communicators are eccentric.
B: Some non-eccentric lecturers are effective teachers but are not good communicators.
C: All good communicators are effective teachers.
D: Some good communicators are eccentric. ✓

Figure 1: Sample from ReClor dataset (Chen et al., 2022)

the first paper which introduced benchmarks on neural models, instead of linear models, and thus there is a limited variety of available benchmarks on the performance of various state-of-the-art transformer models fine-tuned for legal judgement prediction.

Our approach is two-fold: while attempting to devise a model which incorporates semantic relations and logical reasoning, we also implement baselines of existing state-of-the-art models upon the task of multi-label classification for the ECHR dataset ourselves. This is to account for deviations in hyper-parameters such the learning rate, and provide an accurate analysis the impact of each design choice and choice of hyper-parameters upon performance, and especially because the variations in performance for baselines that are presented are extremely subtle for the multi-classification task. for the multi. As a demonstrative case, I adapt the approach of adding a task-specific layer on top of a pretrained model, trained jointly by fine-tuning on the target dataset, and with a softmax layer for prediction.

While relying on a similar conceptual framework as (Chen et al., 2022) to model the logical relations between sequences within the text as a graph, we modify the details of the implementation significantly based on the differences in the model objective and the dataset.

1.0.1 Preprocessing

Obtaining a graph representation of a corpus entails a preprocessing step involving text segmentation. In (Chen et al., 2022), this process is conducted through obtaining the top k n-grams as key phrase nodes (KPH) and relying on a simple chunking algorithm, which segments the texts on conjunctions (e.g. "however", "because") and on any punctuation marks. This phase of the model is not of significance in the case of the ReClor dataset, as demonstrated in the figure1 above, which consists of simple sentence structures and around 3-4 sentences for each sample; yet, in the ECHR dataset, which includes verbose sentences, complex dependencies, and various named entities, the performance of the model is highly dependent on this preprocessing stage. Relying on a simple algorithm as the case mentioned above yields extremely erroneous parses, such as parsing between noun phrases (e.g., chunking on the "," which delimits different components of the same noun phrase in "...from the .... at Leningradskiy railway station in Moskow, L., S., and I., arrested..." (Chalkidis et al., 2019)) which significantly hinders the encapsulation of semantic and conceptual meaning within the graph representation of the text.

Thus, we implement a more sophisticated and thorough dependency parse which both utilizes existing transformer-based models for co-reference and entity extraction, while creating a rule-based algorithm reliant on sentence parsing and word relations to extract phrases in the form of "subject", "action" "object" and "context", each of which are represented as directional nodes (in the order presented) with respect to the graph representation 2. Conjunctions or relations between phrases (units of "subject", "action" "object" and "context" clusters) are represented as fully-connected sub-graphs (through bi-directional edges between all constituent nodes) within the graph, and phrases with the same entities mentioned, which were retained through the coreference resolution and NER indentification step, are represented through bi-directional edges between them.

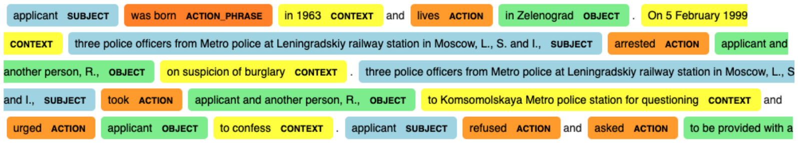

applicant SUBJECT
was born ACTION_PHRASE
in 1963 CONTEXT
and
lives ACTION
in Zelenograd OBJECT
.
On 5 February 1999 CONTEXT
three police officers from Metro police at Leningradskiy railway station in Moscow, L., S. and I., SUBJECT
arrested ACTION
applicant and another person, R., OBJECT
on suspicion of burglary CONTEXT
.
three police officers from Metro police at Leningradskiy railway station in Moscow, L., S and I., SUBJECT
took ACTION
applicant and another person, R., OBJECT
to Komsomolskaya Metro police station for questioning CONTEXT
and
urged ACTION
applicant OBJECT
to confess CONTEXT
.
applicant SUBJECT
refused ACTION
and
asked ACTION
to be provided with a

Figure 2: Phrase Extraction Demonstration

2 Experiments

2.0.1 Graph Construction

After obtaining these segments and their initial embeddings, based on a pretrained token embedding model, for the phrases which are identical, we simply average the embeddings and consider them to be the same node within the graph. We then feed each component of every cluster into an bidirectional gated recurrent unit (BiGRU) to obtain sequence level embeddings, which are the respective nodes of the graph.

As presented in the original graph attention network paper (Veličković et al., 2018), we update the graph network's representations as $\mathbf{h} = \left\{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\right\}, \vec{h}_i \in \mathbb{R}^F$, where N corresponds to the number of nodes in the graph (in our case the number of phrases total), and F the number of features in each node, which corresponds to the sequence embeddings. We perform masked attention, computing e_{ij}, the attention coefficient, for only the first-order nodes, thus nodes $j \in \mathcal{N}_i$ represents node j in the first order neighbourhood of node i in the graph, and

$$e_{ij} = a\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j\right)$$

which communicates the pertinence of node j's features to node i. Coefficients are normalized across all j through a softmax function: $\alpha_{ij} = softmax_j\left(e_{ij}\right) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}$. As the parameters are consistent with the original paper, we simply present the expression for the computation of the coefficient by the attention mechanism here:

$$\alpha_{ij} = \frac{\exp\left(LeakyReLU\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(LeakyReLU\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k\right]\right)\right)}$$

where a is a single-layer feedforward neural network with weight vector $a \in \mathbb{R}^{2F'}$ as the $\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j$ represents the concatenation of $\mathbf{W}\vec{h}_i$ and $\mathbf{W}\vec{h}_k$. The linear combination of the normalized attention coefficients serve as the output features for every node, and thus we get the update rule for the network, from which we extract the last layer for downstream prediction:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N_i} \alpha_{ij} W h_j^{(l)}\right)$$

In line with the methodology of (Chen et al., 2022), we retain the nodes which correspond to the phrases that belong to a given sentence and align them to the sentence embeddings (the [CLS] token, in the case of BERT-based word embeddings), then $\tilde{H}_E = BiGRU\left(H_E + H_{sent}\right) \in \mathbb{R}^{l \times d}$. Finally, we feed the sentence embeddings into an attention layer, emulating the sentence-level attention as in the hierarchical model proposed by (Yang et al., 2016), from which we use a softmax function to get the predictions.

Data. The data that is be used is the ECHR dataset (Chalkidis et al., 2019), which is a set of 11.5k cases from the ECHR's public database. In the dataset, there is a list of facts that have been extracted from the case description, which are mapped to article(s) of the Convention that have been violated, in case of any. The training and development sets are balanced (contain equal number of cases with and without violations), and the test set contains 66% more cases with violations (which is the approximate rate of cases with violations in the database).

	P	R	F1
(Haas and Skreta, 2022)			
LEGAL-BERT	64.8	**59.7**	**62.1**
(Chalkidis et al., 2019)			
HAN	65.0	55.5	59.9
HIER-BERT	**65.9**	55.1	60.0
Jun (2024)			
RoBERTA	63.5	57.0	60.1
XLM-RoBERTA$_{LARGE}$	**65.9**	43.3	52.2
GLJ-BERT$_{CASED}$	65.5	55.2	57.3
GLJ- ROBERTA$_{BASE}$	65.8	55.7	59.0

Figure 3: Results on Experiments for Multi-Classification

Evaluation Method As the shortcoming in performance for existing models is in the multi-label classification task, evaluation is performed on the multi-label classification task.

In line with the evaluation method in Chalkidis et al. (2019), in which the dataset was presented, evaluation is performed on the micro-averaged precision (P), recall (P) and F1 for the multi-label classification task. For the multi-label classification task, LWAN (multi-label violation prediction), BOW-SVM, BIGRU-ATT, HAN, and HIER-BERT (the proposed model) are used in the aforementioned paper, of which HIER-BERT and HAN perform best overall. As of these models, HIER-BERT is the only neural model, yet it suffers from wrongly assigned attention scores due to fact-level attention.

Experimental Details. In line with the aforementioned approach, my experiments were two-part. Firstly, adapting the code and approach presented in the papers by Chalkidis et al. (2019) and Haas and Skreta (2022), I trained the baselines on the first 512 tokens given the word token limit. Furthermore, as the experimental results were consistent with the observation that a high learning rate (of $\alpha = 1\dot{1}0^{-3}$) resulted in a divergence of the training loss (Haas and Skreta, 2022), I also used a learning rate of $\alpha = 2\dot{1}0^{-5}$, and specifically created a baseline with respect to the LEGAL-XLM-RoBERTA$_{LARGE}$, which is a multilingual model pretrained on legal data with the purpose of fine-tuning on downstream tasks.

On the other hand, as previously mentioned, with the intention of attempting to reduce the size of the corpus within the preprocessing stage without the removal of pertinent information. This was done through fine-tuning on a pretrained NER model, (Kalamkar et al., 2022) and using transformer based dependency parsing for co-reference resolution as well as span extractions. Given that this entailed both methods that were manually tailored to the particularities of this dataset as well as an ensemble of pretrained models for the sub-tasks in the preprocessing stage, it required an extremely large amount of time to preprocess the data, given that we could only process the cases one at a time based on resource limitations.

However, unlike the issues regarding training that were faced in other papers which addressed the same dataset, due to the limitation of tokens to 512 tokens per case embedding, as the initial embeddings could be obtained in batches without consideration as to the unrepresented dependencies, as these would be embedded in the graph attention layer of the model.

Results and Discussion. Using ROBERTA-BASE, our model was able to obtain results that are only 0.1-0.2 percent lower than that of state-of-the-art models that are pretrained on large amounts of legal corpa. This is likely attributed to both how the model is able to capture the entire sequence of the corpus and its long-term dependencies, unlike other models which truncate the text. Also, the resolution of co-references explicitly in the pre-processing step, which was necessary for efficient graph-embeddings, compacted the text, which likely improved performance. However, the resource limitations, which entailed that we were only able to process a single case embedding at once, entailed a highly volatile gradient, which was likely a detractor for performance.

With respect to the baseline models tested, as shown in the results below, the performance of XLM-RoBERTA$_{LARGE}$ was equivalent to that of HIER-BERT in terms of precision. This is reasonable, given that both are pre-trained on large amounts of legal documents, therefore show similar performance. However, they clearly demonstrate performance inferior to LEGAL-BERT in terms of recall and F1 scores.

Conclusion and Future Work. Based on the observations from this research, it was evident that the semantic parsing of text was a direct arbiter of performance- yet there was clearly a lack of papers which addressed such issues in the domain relevant to this paper. Consequently, we project that an end-to-end model which also jointly trains on the semantic parsing of text, to learn the phrasal embeddings from the text instead of it being an input parameter, will likely yield high-performance.

Given that legal judgment prediction is a domain which is yet to be fully explored through natural language processing, given that its particularities are directly inter-related to the current limitations and shortcomings in natural language processing methods, we anticipate the application of models which attempt to address logical reasoning into the domain will be a fruitful endeavor for both realms.

References

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Annual Meeting of the Association for Computational Linguistics*.

Jialin Chen, Zhuosheng Zhang, and Hai Zhao. 2022. Modeling hierarchical reasoning chains by linking discourse units and key phrases for reading comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1467–1479, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lukas Haas and Michal Skreta. 2022. From roberta to alexa: Automated legal expert arbitrator for neural legal judgment prediction.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. *ArXiv*, abs/2103.14349.

Yinya Huang, Lemao Liu, Kun Xu, Meng Fang, Liang Lin, and Xiaodan Liang. 2023. Discourse-aware graph networks for textual logical reasoning.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mann Khatri, Mirza Yusuf, Yaman Kumar, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2023. Exploring graph neural networks for indian legal judgment prediction.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *ArXiv*, abs/2004.08994.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Fact-driven logical reasoning. *ArXiv*, abs/2105.10334.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. 2021. Logic-driven context extension and data augmentation for logical reasoning of text. *ArXiv*, abs/2105.03659.

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, San Diego, California. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.