3

Identification of Rhetorical Roles of Sentences in Indian Legal Judgments

Paheli BHATTACHARYA [a,1,2], Shounak PAUL [a,1], Kripabandhu GHOSH [b],
Saptarshi GHOSH [a] and Adam WYNER [c]

[a] *Indian Institute of Technology Kharagpur, India*
[b] *Tata Research Development and Design Centre (TRDDC) Pune, India*
[c] *Swansea University, United Kingdom*

Abstract. Automatically understanding the rhetorical roles of sentences in a legal case judgement is an important problem to solve, since it can help in several downstream tasks like summarization of legal judgments, legal search, and so on. The task is challenging since legal case documents are usually not well-structured, and these rhetorical roles may be subjective (as evident from variation of opinions between legal experts). In this paper, we address this task for judgments from the Supreme Court of India. We label sentences in 50 documents using multiple human annotators, and perform an extensive analysis of the human-assigned labels. We also attempt automatic identification of the rhetorical roles of sentences. While prior approaches towards this task used Conditional Random Fields over manually handcrafted features, we explore the use of deep neural models which do not require hand-crafting of features. Experiments show that neural models perform much better in this task than baseline methods which use handcrafted features.

Keywords. Semantic Segmentation, Rhetorical Roles, Legal Case Documents, Deep Learning, BiLSTM

1. Introduction

Rhetorical role labelling of sentences in a legal document refers to understanding what semantic function a sentence is associated with, such as facts of the case, arguments of the parties, the final judgement of the court, and so on. Identifying the rhetorical roles of sentences in a legal case document can help in a variety of downstream tasks like semantic search [1], summarization [2,3], case law analysis [4], and so on. However, legal case documents are usually not well structured [5,6], and various themes often interleave with each other. For instance, the reason behind the judgment (Ratio of the decision) often interleaves with Precedents and Statutes. Hence it sometimes becomes difficult even for human experts to understand the intricate differences between the rhetorical roles. Hence, *automating* the identification of these rhetorical roles is a challenging task.

For supervised machine learning of the roles, it is important to develop a high quality gold standard corpus, capturing the rhetorical roles of sentences as accurately as possible. Different approaches for the task have constructed their own set of annotated doc-

---

[1] Equal contribution by the first and second authors.
[2] Corresponding Author: Paheli Bhattacharya; Email: paheli.cse.iitkgp@gmail.com

uments [1, 2, 4], but do not report an extensive analysis on the annotation process. Apart from Inter-Annotator Agreement (IAA) scores, it is useful to understand issues such as the amount of subjectivity associated to the labels. In this paper, we perform a systematic annotation study and an extensive inter-annotator study. We show that even legal experts find it difficult to distinguish some specific pairs of labels, thus showing that some subjectivity is inherent in these labels.

Prior attempts to automate the identification of rhetorical roles of sentences in legal documents [2–4] rely on hand-crafted features (see Section 2 for details) such as linguistic cue phrases indicative of a particular rhetorical role [2, 3, 7], the sequential arrangement of labels [2], and so on. Some of these features, e.g., indicator cue phrases, are *largely dependent on legal-expert knowledge* which is expensive to obtain. Also, the hand-crafted features developed in the prior works are often specific to one or a few domains/categories (e.g., Cyber crime and Trade secrets in [4]). It has not been explored whether one can devise a set of features that works for documents across domains.

Recently developed deep learning, neural network models do not require hand-engineering features, but are able to automatically learn the features, given sufficient amounts of training data. Additionally, such models perform better in tasks like classification than methods using hand-crafted features.

In this paper, we explore two neural network models to automatically identify the rhetorical roles of sentences in legal documents – (i) a Hierarchical BiLSTM model, and (ii) a Hierarchical BiLSTM-CRF model. Similar models have been used in the medical domain [8], but to our knowledge, this work is the first to use them in the legal domain. We use these models for supervised classification across *seven rhetorical labels* (classes) and over documents from *five different legal domains*. The Hierarchical BiLSTM-CRF model achieves a very good performance (Macro F-score in the range $[0.8 - 0.9]$), outperforming baseline methods that use hand-crafted features. We also analyse the rhetorical roles predicted by our model, and find that the subjectivity between certain pairs of labels (e.g., Ratio vs. Precedent) that is present among the human annotators is also reflected in the predictions by the algorithm.

This is the first paper on identifying rhetorical roles of sentences in legal documents that brings together (i) an extensive annotation study, and (ii) deep learning models for automating the task. [3]

2. Related Work

In this section we discuss prior work about annotation, automatic rhetorical labelling, and applications of deep learning in the legal domain.

Automatic labelling of the rhetorical role of sentences relies heavily on manual annotation. While papers that aim to automate the task of semantic labelling also perform an annotation analysis [4,5], other works focus on the process of annotation – developing a manual/set of rules for annotation, inter-annotator studies, curation of a gold standard corpus, and so on. TEMIS, a corpus of 504 sentences, that were annotated both syntactically and semantically, was developed in [9]. An in-depth annotation study and curation of a gold standard corpus for the task of sentence labelling can be found in [10], where

---

[3]The dataset and implementations of the proposed neural model are available at *https://github.com/Law-AI/semantic-segmentation*.

assessor agreement was low for labels like Facts and Reasoning Outcomes. Towards automating the annotation task, [11] discusses an initial methodology using NLP tools on 47 criminal cases drawn from the California Supreme Court and State Court of Appeals.

There have been several prior attempts towards automatically identifying rhetorical roles of sentences in legal documents. Initial experiments for understanding the rhetorical/thematic roles in court case documents/judgements/case laws were developed as a part of achieving the broader goal of summarizing these documents [2, 3, 12]. For instance, Saravanan et al. [2] used Conditional Random Fields (CRF) [13] for the task on 7 rhetorical roles. Segmenting a document into functional (Introduction, Background, Analysis and Footnotes) and issue-specific parts (Analysis and Conclusion) was looked into by [4] on U.S. court documents using CRF with handcrafted features. A method for identification of factual and non-factual sentences was developed in [1] using fastText classifier. In another line of work, Walker et al [14] compared use of rule-based scripts (that require much lesser amount of training data) with Machine Learning approaches for the rhetorical role identification task.

Almost all prior attempts towards automatic identification of rhetorical roles in the legal domain have used handcrafted features. In contrast, this paper uses Deep Learning models for this task, where no handcrafted features are needed. Deep Learning (DL) methods are increasingly being applied in the legal domain, e.g., classification of factual and non-factual sentences in a legal document [1], crime classification [15, 16], summarization [6, 17] and other tasks. But, to our knowledge, DL methods have not yet been applied to the task of automatically identifying rhetorical roles of sentences in legal documents.

3. Dataset

In this paper, we consider legal judgments from the Supreme Court of India, crawled from the website of Thomson Reuters Westlaw India (`http://www.westlawindia.com`)[4]. We crawled $53,210$ documents in total. Westlaw assigns each document a legal domain, such as 'Criminal', 'Constitutional', etc. We calculated the frequency of these domains, chose the top 5 domains and randomly sampled 50 documents from these 5 domains in proportion to their frequencies. Thus we have the following set of 50 documents from 5 domains – (i) Criminal – 16 documents (ii) Land and property – 10 documents (iii) Constitutional– 9 documents (iv) Labour and Industrial – 8 documents (v) Intellectual Property Rights – 7 documents. All experiments reported in this paper are performed on these 50 case documents.

4. Annotation Details

In this section we shall describe our annotation study, covering the rhetorical roles / semantic labels we consider in this work, the annotation procedure, and finally, analysis of inter-annotator agreement.

---

[4]We use only the publicly available full text judgement. All other proprietary information had been removed before performing the experiments.

4.1. Annotation Labels / Rhetorical Roles

Our annotators were three senior Law students from the Rajiv Gandhi School of Intellectual Property Law, India (*http://www.iitkgp.ac.in/department/IP*). Based on discussions with the annotators, we consider the following seven (7) rhetorical roles in our work.

1. **Facts (abbreviated as FAC)**: This refers to the chronology of events that led to filing the case, and how the case evolved over time in the legal system (e.g., First Information Report at a police station, filing an appeal to the Magistrate, etc.)
2. **Ruling by Lower Court (RLC)**: Since we are considering Supreme Court case documents, there were some judgements given by the lower courts (Trial Court, High Court) based on which the present appeal was made (to the Supreme Court). The verdict of the lower Court and the ratio behind the judgement by the lower Court was annotated with this label.
3. **Argument (ARG)**: The Court's discussion on the law that is applicable to the set of proven facts by weighing the arguments of the contending parties.
4. **Statute (STA)**: Established laws, which can come from a mixture of sources – Acts , Sections, Articles, Rules, Order, Notices, Notifications, Quotations directly from the bare act, and so on.
5. **Precedent (PRE)**: Prior case documents. Instructions similar to statute citations.
6. **Ratio of the decision (Ratio)**: Application of the law along with reasoning/rationale on the points argued in the case; Reason given for the application of any legal principle to the legal issue.
7. **Ruling by Present Court (RPC)**: Ultimate decision / conclusion of the Court following from the natural / logical outcome of the rationale

4.2. Annotation Process

The annotataors used GATE Teamware tool [18] to annotate the documents, following the methodology of [5, 10]. An annotation manual was developed in discussion with the annotators, containing descriptions and example sentences for each rhetorical role, along with other instructions (e.g., a label should be assigned to a full sentence and not a part of it, a sentence should have only one label, etc.). Initially, each annotator was asked to annotate 5 documents independently, i.e., without consulting each other. Then we had a joint discussion with all the annotators to resolve any issues, and refined the manual if necessary. This process was followed iteratively for annotation of the 50 documents.

4.3. Analysis of Inter-Annotator Agreement and Curation of Gold Standard

We compute the Inter-annotator Agreement (IAA) for the annotation task as follows.

IAA measure: As noted in [10], aggregated pairwise Precision, Recall and F-measure are more suitable measures for IAA than measures like Kappa. Following the same line, we compute these pairwise IAA measures using GATE's Annotation Diff tool.[5] Since we have three annotators (A_1, A_2 and A_3), we compute three sets of pairwise IAA (A_1, A_2), (A_2, A_3), (A_1, A_3), and then take the average of the three sets. We briefly define the metrics below.

GATE maintains three counts based on the extent to which two annotators' labels match. The three counts are as follows – **(1) Correct:** If for a sentence, the two annotators

---

[5]https://gate.ac.uk/sale/tao/splitch10.html

Table 1. Average inter-annotator agreement of the 3 annotators in terms of F-score as measured by GATE tool

Labels →	ARG	FAC	PRE	Ratio	RLC	RPC	STA
Strict	0.692	0.716	0.654	0.677	0.74	0.654	0.857
Lenient	0.953	0.934	0.878	0.908	0.925	0.968	0.967
Average	0.823	0.817	0.814	0.821	0.819	0.798	0.898

mark exactly the same span of text (covering all the words and punctuation marks) with the same label, then this is considered a Correct match. **(2) Partial:** If for a sentence, the two annotators mark the same label, but a different span of text (e.g., leaving few words or punctuation marks), then this sentence is considered a partial match. **(3) Missing and Spurious:** If for a sentence, the two annotators mark different labels, they are called *missing* or *spurious* (both terms used interchangeably). Based on the above definitions, Precision, Recall and F-score are calculated as follows:

$Precision = (Correct + 0.5 \times Partial)/(Correct + Spurious + Partial)$
$Recall = (Correct + 0.5 \times Partial)/(Correct + Missing + Partial)$
$Fscore = ((\beta^2 + 1) \times Precision \times Recall)/((\beta^2 \times Precision) + Recall)$

where β is the weighting of Precision vs. Recall. We use the default value of 1, meaning that both are weighed equally. For each of the Precision, Recall and F-score measures, GATE computes three variants as follows – **(1) Strict measure:** considers all partial matches as *incorrect* (spurious), **(2) Lenient measure:** considers all partial matches as *correct*, and **(3) Average measure:** average of the strict and lenient measures.

Analysis of F-scores: We primarily report the F-scores, since they combine both the Precision and Recall scores. The F-score IAA values computed by using GATE's Annotation Diff tool are presented in Table 1. As is expected, the strict scores are low and the lenient scores are quite high. This is due to differences in how different annotators use the graphical interface of the GATE tool. For instance, one of the annotators may have mistakenly excluded the full-stop (end of sentence marker) in a sentence while marking the label, while the other annotator included the full-stop.[6] The lenient method does not take into account these errors while the strict measure does.

Table 1 reports the IAA (F-score values) for each rhetorical role individually. In terms of strict scores, we observe Statute, Ruling by Lower Court and Facts have a high agreement whereas the scores are lower for Precedent and Ruling by Present Court. But in terms of lenient scores, all labels show high IAA of over 0.85. These IAA scores are comparable with what has been reported in similar prior studies [10].

Analysis of sentence-level agreement: To understand in more detail where the annotators tend to disagree, we perform a *sentence-level agreement study*. We construct an *agreement matrix C* (whose rows and columns are the labels) for two annotators A_x and A_y. An entry $C[i][j]$ of this matrix denotes the number of sentences which Annotator A_x labeled as L_i, but Annotator A_y labeled the *same* sentences as label L_j. Table 2 shows this agreement matrix for the annotator pair (A_2, A_3) who have the *lowest IAA* (as reported by GATE). Similar tables for the annotator pairs (A_1, A_2) and (A_1, A_3) are given in the Supplementary Information accompanying this paper.[7]

---

[6]Though clear instructions were given to include the end of sentence marker in the label, the annotators committed this mistake while marking some of the sentences.

[7]Supplementary Information: `http://cse.iitkgp.ac.in/~saptarshi/docs/Bhattacharya-et-al-JURIX19-SuppleInfo.pdf`

Table 2. Table showing the the sentence level agreement between the two annotators (A_2, A_3) who have the lowest IAA (0.79, as measured by GATE)

$A_2 \downarrow A_3 \rightarrow$	FAC	ARG	PRE	STA	Ratio	RLC	RPC
FAC	**2154**	5	0	3	<u>40</u>	8	0
ARG	<u>17</u>	**822**	16	1	0	0	0
PRE	0	11	**1425**	0	<u>47</u>	0	0
STA	0	0	0	**635**	12	2	0
Ratio	4	13	4	5	**3499**	1	0
RLC	<u>47</u>	1	0	0	<u>25</u>	**294**	0
RPC	6	0	0	0	<u>21</u>	0	**262**

The high values in the diagonal elements indicate that the annotators have a high overall agreement in general. Among the non-diagonal elements, we see relatively high values (signifying some disagreement or subjectivity) for some label-pairs. For instance, there is subjectivity among the label pairs (PRE, Ratio), (FAC, Ratio), (RLC, Ratio) and (RPC, Ratio), since the reason behind the final judgement (Ratio) depends on the facts (FAC), as well as judgements in prior cases (PRE) and the Ruling in the lower courts (RLC). There is also a tendency of annotators to differ between the labels (ARG, FAC) because framing the arguments relies on the facts of the case.

Analysis of agreement across domains: Since we have documents from five domains of law, we checked the average IAA F-score for the labels across each domain. We found that inter-annotator agreement is uniform across different domains. Detailed results can be found in the Supplementary Information.

Curation of the gold standard: The gold standard dataset was curated as follows: For a particular sentence, we took a *majority voting* of the labels given by the 3 annotators. There was a clear majority verdict regarding the label (rhetorical role) of each sentence. We use this annotated dataset in our experiments to automate the task of assigning semantic roles to sentences. Further statistics of the dataset are given in the Supplementary Information.

5. Methods for automatically identifying rhetorical roles

Now we describe our efforts towards automating the task of identifying rhetorical roles of sentences in a legal document. We treat this problem as a 7-class sequence labeling problem, where supervised Machine Learning models are used to predict one label (rhetorical role) for every sentence in a document.

Pre-processing the documents: Each document was split into sentences using the SpaCy tool (`https://spacy.io/`). Splitting a legal document into sentences is challenging due to frequent presence of abbreviations [19]. We observed SpaCy to do a reasonably good splitting (accuracy close to 90%), which agrees with observations in prior works [1]. There were 9,380 sentences in total in these 50 documents, as identified by SpaCy. Each such sentence was considered a unit for which one label (out of the seven rhetorical roles) is to be predicted.

Baseline: CRF with handcrafted features: As stated in Section 2, this is the approach adopted in most prior works. Each document is treated as a sequence of sentences. Some dependencies exist in the corresponding sequence of labels; e.g., RLC usually follow FAC, RPC is always the end label, etc. Conditional Random Fields (CRFs) [13] can be

used to model such sequences, since they consider both *emission scores* (probability of a label given the sentence) and *transition scores* (probability of a label given the previous label) while generating the label sequence.

To implement the baseline approaches [2, 4], we represent each sentence as a vector of all features stated in these works – parts-of-speech tags (used in [4]), layout features (used in both [2, 4]), presence of cue phrases (used in [2]), and occurrence of named entities like Supreme Court, High Court in the sentence (used in [2]). The CRF works on these vectors to predict the labels (rhetorical roles).

We consider three baseline approaches: (1) CRF using the features of [2]; (2) CRF using the features of [4]; and (3) CRF using a combination of features from both [2, 4].

Neural model 1: Hierarchical BiLSTM Classifier: We use a hierarchical BiLSTM (Bi-directional Long Short Term Memory) architecture [20] to automatically extract features for identifying the rhetorical roles. This requires us to feed the sequence of sentence embeddings to the BiLSTM, which returns a sequence of feature vectors. The BiLSTM model needs some initialization of the sentence embeddings, with which learning can start. We try two variations of sentence embeddings: (1) We construct sentence embeddings from *randomly initialized* word embeddings using another BiLSTM; and, (2) We used a large set of documents from the same domain to construct *pre-trained sentence embeddings*. Specifically, we used *sent2vec* [21] to construct the sentence embeddings from the set of 53K court case documents that we had collected (see Section 3), *excluding the 50 documents considered for this task*.

Neural model 2: Hierarchical BiLSTM CRF Classifier: The probability scores generated by the above model do not take into account label dependencies, and thus can be regarded as simple *emission scores*. To enrich the model further, we deploy a CRF on top of the Hierarchical BiLSTM architecture. This CRF is fed with the feature vectors generated by the top-level BiLSTM. As described above, we try both variations of sentence embeddings – randomly initialized embeddings, and pre-trained embeddings trained over a large set of legal documents.

6. Results and Analysis

We now compare the performance of the models (stated in the previous section) on the set of 50 manually-annotated documents (described in Section 4.3).

6.1. Experimental setup and evaluation metrics

We perform 5-fold cross validation with the 50 documents, which is a standard way of evaluating Machine Learning models. In each fold, we have 40 documents for training the model, and the other 10 documents for testing the performance of the model. The performance measures reported are averaged over all five folds.

Evaluation metrics: For a particular sentence, the label (rhetorical role) predicted by a model is considered to be correct, if it matches with the label assigned by the majority opinion of the human annotators (see Section 4.3). We use standard metrics for evaluating the performance of algorithms – macro-averaged Precision, Recall and F-score. For *macro-averaged metrics*, we compute these metrics for each class separately, and then take their average (to prevent any bias towards the high-frequency classes).

Table 3. Macro Precision, Recall and F-score of the baseline methods and neural network-based methods. Best performances highlighted in boldface.

Category	Method	Variations	Precision	Recall	F-score
Baselines	Features from [2]	-	0.4138	0.3308	0.4054
(CRF with handcrafted	Features from [4]	-	0.4580	0.4196	0.3250
features)	Features from [4] and [2]	-	0.5070	0.4358	0.4352
	Hier-BiLSTM	Pretrained emb	0.8168	0.7852	0.7968
Neural models		Random initialization	0.5358	0.5254	0.5236
	Hier-BiLSTM-CRF	Pretrained emb	**0.8396**	**0.8098**	**0.8208**
		Random initialization	0.6528	0.5524	0.5784

Table 4. F-score of the Hier-BiLSTM-CRF model, for the different labels, and for each domain of law. The last column indicates the average F-score for each domain. The last row indicates the average F-score for each of the seven labels (rhetorical roles).

	FAC	ARG	Ratio	STA	PRE	RPC	RLC	*Macro Average (across categories)*
Constitutional	0.903	0.659	0.909	0.832	0.904	0.857	0.85	0.845
Labour & Industrial Law	0.776	0.505	0.929	0.423	0.728	0.783	0.681	0.689
Criminal	0.836	0.567	0.945	0.689	0.891	0.917	0.865	0.816
Land & Property	0.847	0.624	0.908	0.841	0.845	0.98	0.778	0.832
Intellectual Property	0.832	0.607	0.927	0.824	0.901	0.964	0.886	0.849
Macro Average (across labels)	0.8388	0.5924	0.9236	0.7218	0.8538	0.9002	0.812	–

6.2. Comparing performances of different models

The comparative results are presented in Table 3. Clearly the neural models perform much better than the baselines, which shows that the latent features learnt by the neural models are better than the hand-crafted features used in prior works [2, 4].

Effect of pre-trained embeddings: Using pretrained embeddings (learned over a large legal corpus) shows a high improvement in performance for both the neural models, as compared to using random initializations for the embeddings. Since deep neural models require lot of data to learn efficiently, it is especially beneficial to use pretrained embeddings learned over large domain-specific data.

Effect of combining CRF with neural model: Hier-BiLSTM-CRF performs only a little better than Hier-BiLSTM (both with pretrained embeddings). This is because legal documents consist of large sequences (average of 200 sentences per document), and we have few such documents; thus the CRF is unable to learn the transition scores well. Hence, there is not much additional benefit in combining CRF with the neural model.

6.3. Detailed analysis of the best performing model (Hier-BiLSTM-CRF)

Table 4 shows the F-score values of the best performing model (Hier-BiLSTM-CRF) for each of the seven labels and the five domains.

Performance on specific labels: From the last row of Table 4, we find that the model performs the best in predicting the Ratio and Ruling by Present Court (RPC). Ratio has the highest fraction of sentences in the corpus (38.63%), and this large amount of training data enabled this label to be predicted well. Ruling by the Present Court, though having less sentences (2.79% of the dataset), always has a fixed position – towards the end of a document. Hence this label could also be identified well.

Table 5. Label agreement matrix for labels assigned by (i) the best performing Hier-BiLSTM-CRF model, and (ii) majority opinion of the human annotators.

Human ↓ Model →	FAC	ARG	Ratio	STA	PRE	RPC	RLC
FAC	**1986**	109	43	28	35	0	18
ARG	265	**455**	49	22	52	0	2
Ratio	129	51	**3334**	33	72	3	2
STA	57	23	47	**461**	55	0	3
PRE	16	46	64	11	**1330**	1	0
RPC	0	0	9	1	3	**231**	18
RLC	33	5	7	0	7	8	**256**

The model performs satisfactorily for all other labels, except 'Arguments' (F-score of 0.5924). The 'Argument' sentences get interleaved with other labels. Additionally, only 9% of the total number of sentences in our corpus contribute to this label. Hence the neural model did not perform well in identifying these sentences.

Performance across Domains: The last column of Table 4 shows how generalizable the model is across the 5 different domains. The model gives consistent performance (F-score in $[0.82 - 0.86]$) across all the domains, except for 'Labour & Industrial law'. This performance is consistent with the inter-annotator agreement scores, where the IAA was low for the domain 'Labour & Industrial law' (see Supplementary Information).

Comparing inter-annotator agreement and annotator-model agreement: We now compare the agreement between the human annotators (IAA), and agreement between the model and the annotators. We create an agreement matrix (Table 5), where the rows represent the human-assigned labels (majority opinion of the annotators), and the columns represent the labels assigned by the model. The *diagonal elements* show the number of sentences for which the model-assigned label matches with the human-assigned label. The *non-diagonal* elements $C[i][j]$ shows the number of sentences where the human-assigned label i does not match the model-assigned label j.

We focus on the non-diagonal elements that have relatively high values. For instance, the model seems to have frequent confusion between the labels Arguments (ARG) and Facts (FAC), and between the labels Ratio, Fact and Precedence (PRE). Comparing Table 5 with the IAA agreement matrix (Table 2), we find that these label-pairs are exactly the ones where the IAA values were also low, i.e., there is sufficient confusion around these label-pairs even among the human annotators. This observation suggests that these rhetorical roles are largely subjective. Hence it is natural that the model will also face some difficulty in identifying these subjective rhetorical roles.

7. Conclusion and Future Work

We show that deep learning models can much better identify rhetorical roles of sentences in legal documents, compared to methods using hand-crafted features. We also perform an extensive annotation study, and analyse the agreement between different human annotators, as well as the agreement of the model with the annotators.

The principal advantage of neural models is that no hand-crafting of features is needed, hence expensive legal expertise is not essential. However, this property also poses difficulties in understanding why exactly a sentence is more likely to be assigned to one rhetorical role than the others. Thus, neural models trade-off explainability/transparency with the cost of hand-crafting features. Deep Learning models can be

used for tasks like identifying rhetorical roles of sentences, if it can be assumed that achieving good performance is more important than transparency.

In future, we plan to check how deep learning models generalize across different jurisdictions, by experimenting on legal documents of other countries.

References

[1] I. Nejadghoii, R. Bougueng and S. Witherspoon, A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases, in: *Proc. JURIX*, 2017.

[2] M. Saravanan, B. Ravindran and S. Raman, Automatic Identification of Rhetorical Roles using Conditional Random Fields for Legal Document Summarization, in: *Proc. International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[3] A. Farzindar and G. Lapalme, Letsum, an Automatic Legal Text Summarizing System, 2004.

[4] J. Savelka and K.D. Ashley, Segmenting U.S. Court Decisions into Functional and Issue Specific Parts, in: *Proc. JURIX*, 2018.

[5] O. Shulayeva, A. Siddharthan and A.Z. Wyner, Recognizing cited facts and principles in legal judgements, *Artificial Intelligence and Law* **25**(1) (2017), 107–126. doi:10.1007/s10506-017-9197-6.

[6] P. Bhattacharya, K. Hiware, S. Rajgaria, N. Pochhi, K. Ghosh and S. Ghosh, A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments, in: *Proc. ECIR*, 2019.

[7] M. Saravanan, Ontology-Based Retrieval and Automatic Summarization of Legal Judgments.

[8] D. Jin and P. Szolovits, Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts, in: *Proc. EMNLP*, 2018.

[9] G. Venturi, Design and development of TEMIS: a syntactically and semantically annotated corpus of italian legislative texts, in: *Proc. Workshop on Semantic Processing of Legal Texts (SPLeT)*, 2012.

[10] A.Z. Wyner, W. Peters and D. Katz, A Case Study on Legal Case Annotation, in: *Proc. JURIX*, 2013.

[11] A. Wyner, Towards annotating and extracting textual legal case elements, *CEUR Workshop Proceedings* **605** (2010), 9–18.

[12] B. Hachey and C. Grover, Extractive summarisation of legal texts, *Artificial Intelligence and Law* **14**(4) (2006), 305–345.

[13] J. Lafferty, A. McCallum and F.C.N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

[14] V.R. Walker, K. Pillaipakkamnatt, A.M. Davidson, M. Linares and D.J. Pesce, Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning, in: *Proc. Workshop on Automated Semantic Analysis of Information in Legal Texts (with ICAIL)*, 2019.

[15] P. Wang, Z. Yang, S. Niu, Y. Zhang, L. Zhang and S. Niu, Modeling dynamic pairwise attention for crime classification over legal articles, in: *Proc. ACM SIGIR*, 2018.

[16] P. Wang, Y. Fan, S. Niu, Z. Yang, Y. Zhang and J. Guo, Hierarchical Matching Network for Crime Classification, in: *Proc. ACM SIGIR*, 2019.

[17] C.-L. Liu and K.-C. Chen, Extracting the Gist of Chinese Judgments of the Supreme Court, in: *Proc. ICAIL*, 2019.

[18] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, in: *Proc. ACL*, 2002.

[19] G. Sanchez, Sentence Boundary Detection in Legal Text, in: *Proc. Natural Legal Language Processing Workshop*, 2019.

[20] A. Graves, S. Fernández and J. Schmidhuber, Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition, in: *Proc. Int'l Conf. on Artificial Neural Networks (ICANN)*, 2005.

[21] M. Pagliardini, P. Gupta and M. Jaggi, Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features, in: *Proc. NAACL*, 2018.