Semi-Supervised Methods for Explainable Legal Prediction

K. Branting
B. Weiss
B. Brown
The MITRE Corporation
McLean, VA, USA
{lbranting,bweiss,bcbrown}@mitre.org

C. Pfeifer
The MITRE Corporation
Ann Arbor, MI, USA
cpfeifer@mitre.org

A. Chakraborty
L. Ferro
M. Pfaff
A. Yeh
The MITRE Corporation
Bedford, MA, USA
{achakraborty,lferro,mpfaff,asy}@mitre.org

ABSTRACT

Legal decision-support systems have the potential to improve access to justice, administrative efficiency, and judicial consistency, but broad adoption of such systems is contingent on development of technologies with low knowledge-engineering, validation, and maintenance costs. This paper describes two approaches to an important form of legal decision support—explainable outcome prediction—that obviate both annotation of an entire decision corpus and manual processing of new cases. The first approach, which uses an Attention Network for prediction and attention weights to highlight salient case text, was shown to be capable of predicting decisions, but attention-weight-based text highlighting did not demonstrably improve human decision speed or accuracy in an evaluation with 61 human subjects. The second approach, termed SCALE (Semi-supervised Case Annotation for Legal Explanations), exploits structural and semantic regularities in case corpora to identify textual patterns that have both predictable relationships to case decisions and explanatory value.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

Artificial Intelligence & Law, Legal Reasoning, Computational Models of Argument, Machine Learning, Human Language Technology

ACM Reference Format:
K. Branting, B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, and A. Yeh. 2019. Semi-Supervised Methods for Explainable Legal Prediction. In *Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19), June 17–21, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3322640.3326723

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICAIL '19, June 17–21, 2019, Montreal, QC, CA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6754-7/19/06.
https://doi.org/10.1145/3322640.3326723

1 INTRODUCTION

Recent advances in Artificial Intelligence (AI) and Human Language Technology (HLT) have created new opportunities to automate routine aspects of case management and adjudication, freeing human experts to focus on aspects of these tasks that most require human judgment and knowledge. An important application of this technology is decision support for routine administrative decision-making and adjudication. Globally, benefits adjudications, resolution of commercial conflicts, criminal defense, and other forms of access to justice are often impeded by the opacity of legal processes, shortages of affordable legal assistance, and growing case backlogs [19]. Effective decision-support systems could potentially improve access to justice for significant numbers of citizens by improving transparency, compensating for lack of affordable human legal assistance, and speeding case decisions.

A particularly simple but useful form of decision support is *explainable legal decision prediction*. For example, benefits applicants could make better-informed decisions if they knew (1) the likelihood of success of an application and (2) the strengths or weaknesses of their application, i.e., the reasons for the predicted likelihood. Similarly, adjudicators and decision processes could be more productive and consistent if the strengths or weaknesses of each application could be automatically identified and presented along with the application itself.

Inherent in explainable outcome prediction systems is a trade-off between *explanation quality* and *development effort*. At one extreme, purely machine-learning-based systems require relatively little development effort but typically have little or no explanatory capability. At the other extreme, systems in which case facts have been represented in manually engineered features and legal rules are represented in executable logic may be capable of generating explanations with considerable fidelity to human explanations but require prohibitively high levels of development effort for application at scale.

A key requirement for explainable decision predictions systems is therefore optimizing the trade-off between explanation quality and development effort, that is, identifying approaches that can produce useful and comprehensible predictions but for which the engineering effort is low enough to permit development, verification, and maintenance at scale. A particularly important requirement for large-scale adoption is the ability to accept free text rather than manually-engineered features as input.

This paper describes two approaches to explainable legal decision prediction that operate on textual inputs. Each system was

Paradigm	K-E artifacts	Execution-time representation	Output	Example
1. text → output			prediction, relevance weights	NFE
2. text → predicates → output	predicate set, rule set		prediction, per-predicate relevance weights	
3. text → features → output	feature set		case-based argumentation	
4. text → features→ predicates → output	feature set, predicate set, rule set		hybrid case/rule-based argumentation	SCALE
5. features → output	feature set	featural case representation	case-based argumentation	HYPO
6. features → output	rule set	featural case representation	rule-base argumentation	
7. features→ predicates → output	feature set, predicate set, rule set	featural case representation	hybrid case/rule-based argumentation	NIHL (Angelic methodology)

Figure 1: Paradigms of explainable decision prediction.

prototyped on a collection of 16,024 World Intellectual Property Organization (WIPO) domain name dispute cases.

The first system, NFE (No Feature Engineering), uses an Attention Network [28] for prediction and highlights salient case text based on attention weights for decision support. NFE was evaluated as a decision aide for attorney and non-attorney subjects to predict the case decisions. While the case decisions themselves were found to be predictable using this approach, attention-weight-based text highlighting was not shown to improve decision speed or accuracy. This negative result motivated a second approach, termed SCALE (Semi-supervised Case Annotation for Legal Explanations), in which the justification structure of a representative set of cases is annotated, and tags for factual and legal findings are propagated to sentences in unannotated cases that share a high degree of similarity to the annotated sentences in a semantic embedding space. This approach exploits the structural and semantic regularities in case corpora to identify fact patterns with predictable relationships to case decisions.

The remainder of this paper is structured as follows. Section 2 discusses the inherent trade-off between explanation quality and development effort. A minimalist approach to feature engineering, NFE, is discussed in Section 3, and Section 4 discusses a semi-supervised approach, SCALE, motivated by the negative results of the evaluation of NFE. The annotation scheme is described in Section 5, and Section 6 discusses how annotations of a representative set of cases are mapped onto similar sentences from other cases in semantic embedding space for use in prediction and decision support. The paper concludes with related and future work.

2 THE TRADE-OFF BETWEEN EXPLANATION QUALITY AND DEVELOPMENT EFFORT

Until the recent rise in popularity of data science and deep learning in AI and law, the primary focus of research in this community was on computational models of argumentation [8]. These models typically operated on manually engineered representations of case facts and produced argument trees or other formal justifications for legal conclusions [27]. Even when the objective was simply prediction, rather than justification, case features were typically assigned to each case manually [20] [2]. In general, the only circumstance under which manual representation of the facts of each new case was obviated was when the user interface queried users for individual feature values [30] [25] or when prediction was based on factors unrelated to the merits of the case, such as the attorneys, judge, cause of action, etc. [33]. When the features closely corresponded to users' conception of relevant case facts, this approach was workable. However, this approach often requires a significant knowledge-engineering effort to identify a cognitively-plausible feature set [1].

Advances in machine learning and corpus-based techniques have made outcome prediction increasingly feasible, even in the absence of domain-specific features. Several projects have demonstrated the feasibility of predicting case outcomes from unprocessed text, either using deep learning techniques or by applying symbolic machine learning techniques to n-grams or other domain-independent lexical or linguistic text features, provided that there are sufficient training examples [3] [13] [32]. However, such systems have very limited inherent explanatory capability. Reducing the opacity of the machine learning algorithms that perform best for many purposes—neural network models, often referred to as Deep Learning—is the focus of very active current research, such as the DARPA "Explainable AI" (XAI) program [18]. However, there has been little XAI work directed at the particular forms of explanation characteristic of legal discourse.[1] Legal justifications and explanations differ from typical XAI applications in that they must make explicit reference to authoritative legal sources to be persuasive.

One way to characterize the dependence of explanation quality on prior engineering effort is by identifying the knowledge-engineering artifacts intermediary between the textual expression of the facts of a case and the output (e.g., a prediction with some degree of explanation or justification). Figure 1 sets forth a notional division of explainable prediction systems. In this figure, the term "feature" is intended to mean a fact pattern of potential legal relevance,[2] whereas a "legal predicate" is a term or concept occurring as an antecedent or consequent of a legal rule or norm.

Paradigm 1 is a pure machine-learning approach requiring no knowledge engineering other than construction of an output-labeled corpus and producing little explanation beyond the prediction itself and the internal model weights associated with that prediction (e.g., attention weights on input subtexts). At the opposite extreme, Paradigm 7 can generate hybrid case-based and rule-based argumentation but requires manually engineered feature, issue, and rule sets and execution-time representation of new cases in terms of that feature set [26] [12].

The objective of the research described in this paper is to develop techniques for explainable prediction that can provide practical tools for improving the efficiency of administrative decision processes without imposing excessive knowledge-engineering, validation, and maintenance costs on agencies. Minimizing up-front engineering costs is important because the factors motivating agencies to consider decision-support systems—decision backlogs resulting from insufficient resources—tend to preclude solutions that require extensive development efforts.

[1] See [35] for a recent exception to this generalization.
[2] The "factors" of [4] are examples of features in this sense.

The domains of particular interest to us include disability benefit claims, immigration petitions, landlord-tenant disputes, and attorney misconduct complaints. The high volumes of cases of these types mean both that large training sets are available and that agencies have an incentive to consider technologies to improve decision processes. However, because of the sensitivity of most of these data sets, we have found it convenient to perform initial research on an open-source data set consisting of 16,024 World Intellectual Property Organization (WIPO) domain-name dispute decisions.[3]

WIPO cases have only two possible decisions: granting or denying the request to transfer a domain name to the Complainant. WIPO cases are consistently segmented into seven sections: Parties, Domain Name, History, Background, Contentions, Findings, and Decision. The Findings section typically consists of four subsections, one for each of the three requirements that Complainants must establish to prevail (*issues*): "confusability", "no legitimate rights or interests", and "bad faith", plus an additional issue—reverse name highjacking—that we do not address. We formed a case-decision corpus in which the facts of each instance consist of the concatenation of the first 5 sections, and the decision is "transferred" or "not transferred." This dataset has a roughly 10-to-1 class skew in favor of "transferred." In addition, we formed a separate corpus for each of the three issues, for a total of 4 datasets used in the experiments below.

The next section describes development and evaluation of NFE, the minimalist approach to explainable decision prediction.

3 A NO-FEATURE-ENGINEERING PARADIGM

The No Feature Engineering (NFE) approach (Paradigm 1) treats legal prediction as a problem of classifying texts that represent case descriptions into decision categories. This approach may be appropriate when there is a corpus of previous case-text/decision pairs—as is typically the case in administrative or judicial forums—and decisions can be viewed as unstructured categories, such as granting or denying a particular benefit or form of relief.[4] Prototypical domains fitting this description involve routine administrative decisions, such as applications for permits, licenses, or benefits, each of which is characterized by relatively circumscribed and predictable fact patterns and a limited range of permissible outcomes. Cases in trial or appellate courts of general jurisdiction are generally ill-suited to this approach because of the wide variety of both facts giving rise to cases and types of decisions.

This shallow approach to decision prediction is not appropriate for completely autonomous processes. Denial of benefits by an automated process, no matter how accurate, raises significant due-process issues, and in any event prediction accuracy is limited in this paradigm by the absence of explicit modeling of legal rules or issues. However, even an imperfect prediction can be useful in high-volume forums for (1) *triage*, e.g., granting benefits without further processing in extremely clear cases or routing cases to particular decision makers based on apparent complexity, and (2) decision support.

[3] http://www.wipo.int/amc/en/domains/decisionsx/index-gtld.html.

[4] Cases in which decisions consist of numerical awards can be modeled as regression problems. For simplicity, we confine the discussion in this paper to categorical classification.

For example, [13] applied three different machine-learning approaches to an earlier version of the WIPO dataset: a Hierarchical Attention Network (HAN) [36], Maximum Entropy (Maxent) classification [9], and a Support Vector Machines (SVM), the latter of which were applied to a token n-gram representation of the case text. The frequency-weighted f1 for all three models was roughly 0.91, in contrast to an f1 of about 0.86 for the majority-class rule. For the HAN, the Matthews Correlation Coefficient (MCC), a standard measure of the improvement of a binary classifier over a majority class rule that ranges from -1 to +1, was about 0.558.

3.1 Explanation in No-Feature-Engineering Systems

Since the NFE approach lacks explicit legal concepts, our approach to decision support focused on identifying the portions of case text that are most predictive of the decision. Our hypothesis was that a decision maker may benefit from having the predictive text identified even in cases in which the decision disagrees with the models prediction. This hypothesis was based on the observation that one of the challenges of decision making is sifting through irrelevant portions of the case record to locate the most important facts.

One approach to predictive-text identification would be to produce a case summary that includes only the most relevant portion of the case facts. This might approximate the actions of a law clerk or other administrative assistant. However, many decision makers may prefer to see the text selected as most relevant in its original textual context. Consistent with this approach, we experimented with various approaches to highlighting important portions of the case text, including pairwise mutual information and logistic-regression model weights. However, attention-network weights [7] have the advantage over these methods in that the weights are specific to each individual case rather than global for the entire corpus. In the evaluation below, we therefore focused on attention-weight-based highlighting.

3.2 Evaluation of Attention-Based Highlighting for Decision Support

Prior work had not established whether simply highlighting the most important portion of a case description based on attention weights can provide significant decision support. We therefore performed an evaluation using an attention network trained on the WIPO cases described above.

A total of 61 participants employed by the MITRE Corporation were recruited for the study: 34 with legal experience and 27 without legal experience, but none familiar with the WIPO domain. Participants were randomly assigned to 1 of 4 conditions:

- Control: Case text only
- Highlighting: Case text plus highlighting
- Precedents: Case text plus positive and negative precedents
- Highlighting and Precedents: Case text plus highlighting and positive and negative precedents

Each participant was asked to decide the issue of "No Rights or Legitimate Interests" (NRLI) in two separate cases and to provide a justification for each prediction. In addition, they were asked to

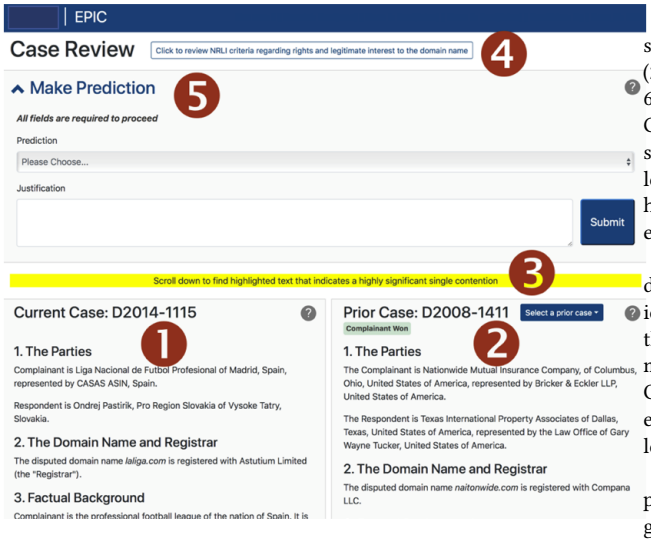

Figure 2: Screenshot of human evaluation

complete a survey to collect demographics and opinions about the experience of the tasks and usability of the interface. Participants were encouraged to complete the study within an hour.

The screenshot in Figure 2 shows the interface as it was presented to participants in the condition with both highlighting and precedents. All participants saw the "Current Case" (**1**) for which they were to predict the NRLI decision. Participants in conditions that included the positive and negative precedents were shown in the "Prior Case" pane (**2**) to the right of the Current Case. Participants were able to toggle between two prior cases, one in which the Complainant won, and one in which the Complainant lost on the NRLI issue. For participants in conditions including highlighting, certain passages of text were highlighted in yellow, as prompted by the yellow bar above the cases (**3**). At any time, participants could refer back to the NRLI criteria (**4**). Finally, participants made their prediction (NRLI is true or false) and wrote a justification for that prediction (**5**). Four dependent variables were assessed for each trial:

- Correctness: Participant prediction (NRLI is true or false) compared to actual case outcome.
- Task Duration: Amount of time to evaluate case, timed from when the case is first presented until the participant submits the prediction with justification.
- Task Difficulty: Measured with Single Ease Question [29] and eight items selected from the Multiple Resources Questionnaire [11].
- User Satisfaction: Measured with System Usability Scale [14].

More participants decided Case 1 correctly (48) than incorrectly (13). However, participants with legal experience were significantly more likely to decide Case 1 correctly (30 out of 34, or 88.2 %) than those without legal experience (18 out of 27, 66.67%), χ^2 (1, N=61) = 4.18, p = .04.

Highlighting had no effect on correctness, but participants were significantly more likely to predict Case 1 correctly with precedents (25 out of 27, or 92.6%) than without precedents (23 out of 34, or 67.6%), χ^2 (1, N=61) = 5.58, p = .02. More participants decided Case 2 incorrectly (53) than correctly (8). In Case 2, there were no statistically significant differences between those with and without legal experience and there were no significant effects of either highlighting or precedents on correctness. Participants with legal experience spent more time on the task for both cases.

The observation that participants were more likely to correctly decide Case 1 with precedents, even though it took them significantly more time to decide than without precedents, suggests that precedents contribute important information to the decision-making process even in a domain, like WIPO, without *stare decisis*. Overall, however, this study did not find statistically significant evidence that providing outcome-relevant case comparison features leads to improved decision making.

An important proviso to these conclusions is that some participants' comments indicate that they mistakenly believed that their goal was to predict the outcome of the entire case—whether the domain name should be transferred—rather than simply the NRLI issue, a confounding factor not evident in the pilot conducted in advance of the main study. This illustrates that non-experts can find deciding even simple administrative issues very challenging and that the design of decision-support systems for such users (e.g., citizens not represented by attorneys familiar with the domain) requires extensive usability analysis. Another proviso is that this study doesn't rule out the possibility that highlighting produced by a different predictive model might be more useful for decision support.

Perhaps the most illuminating comments by participants were that they had difficulty understanding the connection between the highlighted text and the issue that they were supposed to decide. These comments, and the overall results of study, indicate that useful decision support should help the user understand the connection between relevant portions of the case record and the issues and reasoning of the case.

4 SCALE: SEMI-SUPERVISED CASE ANNOTATION FOR LEGAL EXPLANATIONS

The results of the evaluation of NFE suggest that an effective decision support system for prediction explanations must identify not just relevant case text, but fact patterns and issues that connect the text to the predicted outcome. To achieve this capability in the WIPO domain without individually annotating every case, we have developed a novel approach based on exploiting regularities in opinion structures. This approach is based on several observations about the consistency of language across separate cases and within different sections of the same case.

First, we observe that the relatively stereotyped language of administrative case decisions means that statements with similar legal effect in different cases tend to be close to each other in semantic-embedding space [24]. Thus, annotation tags applied to a subset of cases can be mapped to an entire corpus, with accuracy and completeness that depends on the consistency of the case language within the corpus, the typicality of the annotated cases, and the

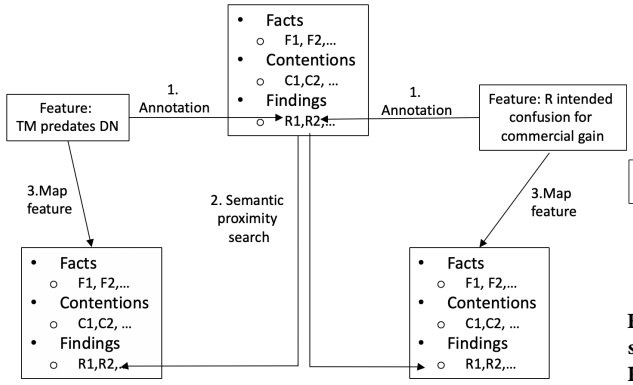

Figure 3: Annotation, clustering, and tag mapping.

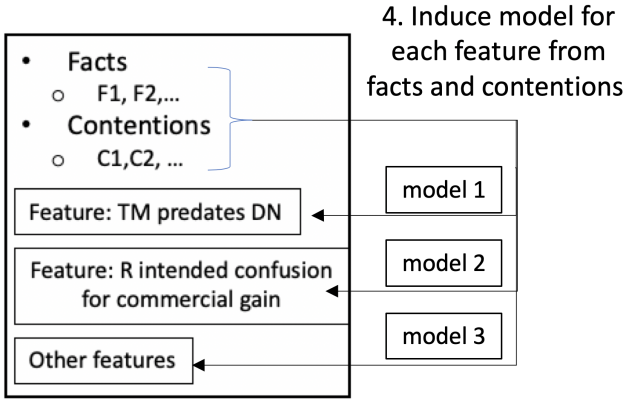

Figure 4: Inducing a model for each feature from facts and contentions.

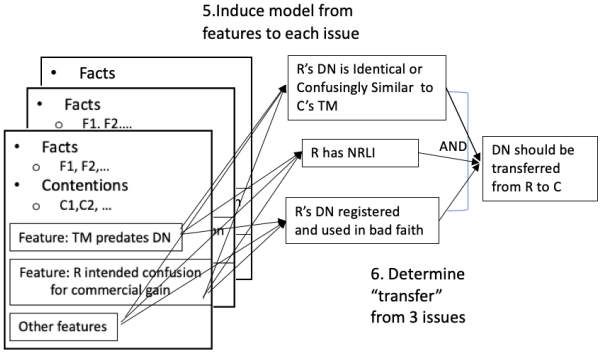

Figure 5: Inducing and applying feature-to-issue models.

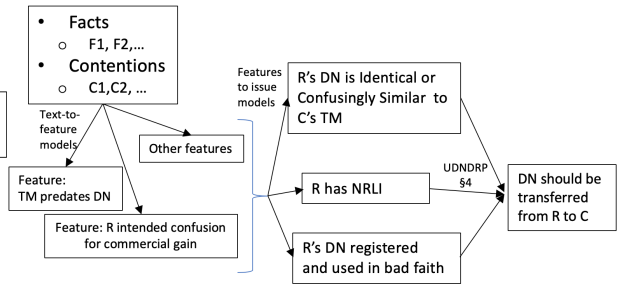

Figure 6: Analysis of a new case. "UDNDRP" represents Section 4 of the Uniform Domain Name Dispute Resolution Policy (https://www.icann.org/resources/pages/policy-2012-02-25-en).

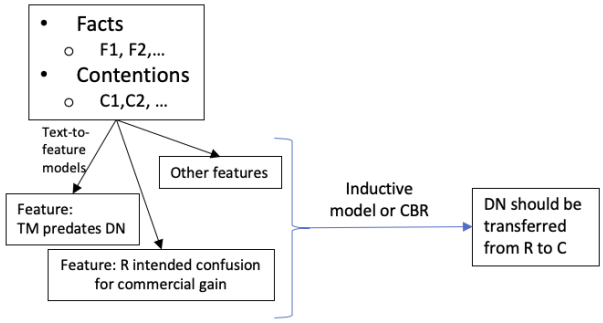

Figure 7: Case analysis directly from features to decision without intermediate issues.

threshold for semantic similarity. Second, most factual-finding or legal-ruling sentences correspond semantically to one or more sentences in the contentions section. This is a manifestation of the inherent property of legal decisions that findings and rulings resolve contentions by parties. Third, the polarity of each finding and rule, that is, which party it supports, depends on sentences in the facts section. Thus, machine-learning techniques developed to predict the overall outcome of the case can be applied as well to predict the polarity of individual findings and rulings.

Finally, in many administrative domains the Findings or Decision sections of cases are subdivided into predictable subsections, each resolving one of the elements ("issues") that a Complainant must establish. In WIPO cases, there are 4 subsections. Considering each subsection separately permits the overall decision to be broken into separate steps, improving comprehensibility.[5] Our key hypothesis is that these document regularities can be exploited to project annotations from a representative set of decisions onto the entire corpus, and that the resulting semi-automated corpus can be exploited to identify the case features that (1) predict the issues on which the decision will turn and the decision itself, (2) justify

[5] At the time of writing, we have not yet completed annotation of each individual issue for all instances in our data set. The experiments described below therefore involve prediction only of the overall outcome of the case without individual issue decisions.

a prediction in terms of the features of the particular case, and (3) identify prior cases whose facts and contentions are most relevant to a given case.

As shown in Figure 3, Step 1 of the SCALE procedure consists of manual annotation of the Findings section of a representative set of cases. All sentences in the corpus in close proximity to a tagged sentence in the semantic-embedding space are identified in Step 2, and in Step 3 the corresponding tags themselves are mapped to all similar sentences from the Findings section of some case. The result of these steps is an annotation of the Findings section of every case in the corpus. These mapped annotations will almost certainly be less accurate than manual annotations; the actual accuracy will depend on the details of both the original annotation and the clustering itself. However, we hypothesize that such mapped annotations can be accurate enough for prediction, triage, and decision support.

Figure 4 shows the next step, inducing a separate model for each feature from the case description, i.e., the Facts and Contentions. In Step 5, shown in Figure 5, feature-to-issue models are induced for each of the high-level issues. Figure 6 illustrates how the outcome of a new case can be predicted by first predicting features from Facts and Contentions, predicting issue-decisions based on features, and finally predicting the "transfer" decision based on the issues. Figure 7 shows the simplified process used in the experiments described below that skips issue prediction and instead goes directly from features to case outcome.

The SCALE approach differs from NFE in it that involves reasoning about case features induced from the case description which can be used to explain and justify a prediction. Our initial implementation involves annotation of 16,024 WIPO decisions.

5 ANNOTATION

A key goal of SCALE is a methodology that permits development of explainable legal prediction systems by agencies that lack the resources to engineer domain-specific feature sets, a process that requires both extensive expertise in the particular legal domain and experience in feature engineering. Instead, SCALE requires only the linguistic skills necessary to annotate the decision portion of representative subset of cases, a much more limited process.

Our annotation schema for WIPO decisions consists of three layers: Argument Elements, Issues, and Factors (sub-issues).[6] Tags are applied to clauses and sentences, as opposed to shorter units such as noun phrases, in order to identify the complete linguistic proposition corresponding to the annotation label. The MITRE Annotation Toolkit (MAT)[7] was used to perform the annotation.

5.1 Argument Elements

Although our approach to predictive-text identification is to leverage the Factual Findings and Legal Findings, the annotation schema is designed to capture the full range of argument elements present in cases. These argument elements are as follows:

(1) Policy
(2) Contention
(3) Factual Finding
(4) Legal Finding
(5) Case Rule
(6) Decision

We have found that with these six argument elements, the majority of sentences within the "Discussion and Findings" and "Decision" sections of WIPO cases can be assigned an argument element label. These argument elements are not specific to WIPO decisions and should be applicable in other domains.

5.2 Issues

Each Argument Element tag is assigned an Issue. The Issue tags include the three required elements that the complainant must establish in order to prevail in a WIPO case. These issues, which are documented in the Uniform Domain Name Dispute Resolution Policy, paragraph 4,[8], form the backbone of every decision:

- (i) ICS: Domain name is Identical or Confusingly Similar to a trademark or service mark in which the complainant has rights
- (ii) NRLI: Respondent has No Rights or Legitimate Interests in respect of the domain name
- (iii) Bad Faith: Domain name has been registered and is being used in Bad Faith

For element (ii), NRLI, although the dispute is typically approached from the point of view of the complainant demonstrating that the respondent has NRLI, it is very often the case that the panel considers the rights or legitimate interests of the complainant and/or the respondent. In that case, RLI is available as an Issue tag.

In addition, the domain name resolution procedure allows for situations in which the complainant abuses the process by filing the complaint in bad faith (CIBF).[9]

The schema thus consists of five Issue tags, plus an Other category:

- ICS
- NRLI
- RLI
- BadFaith
- CIBF
- OTHER

5.3 Factors

In our annotation scheme *factors* are the elements which we hypothesize will prove most useful for explainable legal prediction. The factors and corresponding tags are specific to the WIPO issues. For ICS, the ICANN policy does not explicitly identify specific factors that will be considered by the panel, so our tag set for ICS is derived from factual findings commonly observed in the data, such as CownsTM (Complainant owns Trademark) and TMentire (Trademark is contained in its entirety within the Domain Name). For NRLI/RLI, the policy establishes three factors, and for Bad Faith, four factors. Each of these has a corresponding tag. For example, under NRLI there is PriorBizUse from 4(c)(i) of the policy ("Bona

[6] We are also currently exploring a fourth layer, Evidence, which captures the evidence cited in support of the Factor or Issue. As this exploration is still underway and the Evidence tag set continues to expand, it will not be discussed further here.
[7] http://mat-annotation.sourceforge.net/

[8] https://www.icann.org/resources/pages/policy-2012-02-25-en#4
[9] See 15(e) of the Rules for Uniform Domain Name Dispute Resolution Policy for CIBF, https://www.icann.org/resources/pages/udrp-rules-2015-03-11-en

Case No.	Text	Annotation
D2016-0534	The Complainant must have been aware that the Disputed Domain Name existed when it chose to register its UNIKS trademark	FACTUAL_FINDING-CIBF-RDNH
D2016-0534	in two instances the TURBOFIRE mark has been reproduced in a domain name, utilizing a dash "—" between the "turbo" and "fire" portion of the mark, which the Panel disregards as irrelevant under this element of the Policy	FACTUAL_FINDING-ICS-IrrelevantDiff
D2012-1430	the Respondent clearly is not making any noncommercial or fair use of those domain names	LEGAL_FINDING-NRLI-LegitUse subissue-polarity=negative
D2012-1430	Such use constitutes bad faith under paragraph 4(b)(iv) of the Policy	LEGAL_FINDING-BadFaith-Confusion4CommGain

Figure 8: Four text spans annotated with factual and legal-findings features.

fide business use of Domain Name or demonstrable preparations to do so, prior to notice of the dispute") and under BadFaith there is Confusion4CommGain from 4(b)(iv) of the policy "For commercial gain from confusion with complainant's mark"). The tag set also includes labels for other common factors observed in the data, such as PrimaFacieEst (Prima Facie Case Established).

For CIBF, two factor tags are available: RDNH (Reverse Domain Name Hijacking) and Harass (complaint brought primarily to harass DN holder).

Each level of annotation also has an "Other" option to be used when none of the predefined tags is appropriate, and there is a free-form Comment field which the annotator can use to capture ad hoc labels and enter notes.

5.4 Attributes

A Citation attribute is used to capture the paragraph citation of Policy and Case Rule argument elements. A polarity attribute is used to capture positive/negative values for issues and factors. Figure 8 shows four typical annotations.[10]

6 MAPPING ANNOTATIONS FOR PREDICTION AND EXPLANATION

Linguistic annotation is an expensive and arduous task, but necessary to train and evaluate analytics. From a small set of 25 annotated documents (0.156% of the entire corpus), we were able to project the annotations to similar sentences throughout the entire corpus of documents. This projection was accomplished through the use of word and sentence embeddings to find text that is semantically similar to the annotated text. The projection method is as follows.

First, word embeddings are trained on the tokenized corpus using FastText[10]. This yields one vector per token that captures the semantics of the word through the surrounding context. Next, these word embeddings are used to compute sentence embeddings by averaging the vectors of the words in each sentence for each of the 2.64 million sentences in our corpus. Semantically similar sentences are close to each other in semantic-embedding space. A notable limitation of this approach is that sentences that are lexically very similar but that have opposite polarity are often very close in this embedding space. An example is simple negation via "not", for example "the panel finds that it was properly constituted" and "the panel finds that it was not properly constituted" differ by a single word but have opposite legal effects. This does not impact our current annotation as the inventory is primarily focused on capturing argument elements, leaving polarity determination to a subsequent processing step. The sentences are then clustered into 512 clusters by cosine similarity in embedding space. The clusters establish neighborhoods of semantically similar sentences.

Once the word embeddings have been trained, embeddings for the annotation spans of text are trained using the same method as was used to compute sentence embeddings. While the annotation spans are not strictly sentences, the sentence embedding method can be used to compute embeddings of arbitrary spans of text.

Once the word embeddings, corpus sentence embeddings, annotation span embeddings and clusters have been computed, the tags can be projected. For each annotation label of interest for the specific experiment, we retrieve the top 10,000 sentences in the corpus ranked by cosine similarity to the annotated spans. Then, the annotation label is projected to each cluster associated with each retrieved sentence.

For these prediction tasks, we do not use the words of the sentences. Instead, we use the cluster label of each sentence in the document. The sentences are selected according to task-specific criteria. XGBoost[17], an efficient implementation of the gradient boosting algorithm, is used in all prediction tasks in this work. These are preliminary results, and we continue to iterate to improve the outcomes.

As the transfer decision labels are highly skewed, 91% transfer (14,591 cases), 9% nontransfer (1,407 cases), we do not create a dedicated test set. Instead, we opt for 10 random test/train splits and report the average area under the curve (AUC), and per class precision, recall and F1 score aggregated over the 10 trials. A summary of the results is given in Table 1.

The results set forth below are preliminary results intended to demonstrate the feasibility of the SCALE approach. We expect that expanding the set of annotated documents beyond its small initial size (25) should improve performance.

6.1 Predicting Decisions from Mapped Tags

The accuracy of mapped tags as predictive features depends on both the annotation conventions and the details of the clustering. An initial evaluation of adequacy and correctness of these initial two steps can be performed by determining whether the mapped tags can be used to predict case decisions. If the tags are capturing the actual decision, then a high degree of accuracy should be achievable by training a model that predicts overall case decisions, or decisions for individual issues, from the mapped tags.

This experiment used the tag projection method described above and retrieved sentences based on all annotation types. This method selected 1.8M sentences out of the total corpus of 2.6M.

Predicting overall case outcome with the annotated data gave strong results with an average AUC of 80.8% and a standard deviation of 0.01. The transfer class, the majority class in this dataset, earned a 91.9% F1 (97.3% precision and 87.1% recall). The non-transfer class was lower with a 48.2% F1 (35.7% precision, 74.5%

[10]We plan to make this annotated corpus available to researchers in 2019 at https://github.com/mitre.

Prediction Task	Avg. AUC	Std. Dev.	Positive/Transfer Precision	Recall	F1	Negative/Non-Transfer Precision	Recall	F1
6.1 Predict Decisions from Mapped Tags	80.8%	0.01	97.3%	87.1%	91.9%	35.7%	74.5%	48.2%
6.2 Predict Decisions from Factual Findings Tags	89.6%	0.008	98.8%	90.2%	94.3%	46.7%	89%	61.2%
6.3 Predict "(ICS)-…" Finding from Case Text	60.1%		38.3%	42.3%	40.2%	80.9%	78.2%	79.5%
6.3 Predict "(NRLI)-…" Finding from Case Text	62.9%		31.7%	54.6%	40.1%	86.5%	71.1%	78.1%
6.3 Predict "Bad Faith-…" Finding from Case Text	63.5%		43.9%	48.9%	46.3%	81.5%	78.2%	79.8%

Table 1: Summary of Prediction Results

recall). This experiment indicates that tags mapped from a modest set of 25 annotated cases (0.156 percent of the entire corpus) are sufficient to express the decisions in the Findings section.

6.2 Predicting Decisions from Factual Findings Tags

The next experiment involved restricting the prediction to just those tags that represent factual findings. The purpose of this experiment is to determine whether the factual findings are sufficient to determine the outcome of individual issues and of the case as a whole. This is important because it established that if factual findings can be predicted from the case text (e.g., from all the sections preceding the panel's discussion and findings) then the outcome could be predicted from these findings. This experiment used the tag projection method described in section 6, and retrieved sentences only of factual finding annotation types. This method selected 1.3M sentences out of the total corpus of 2.6M sentences.

Predicting overall case outcome with the annotated data gave strong results with an average AUC of 89.6% and a standard deviation of 0.008. The transfer class, the majority class in this dataset, earned a 94.3% F1 (98.8% precision, 90.2% recall). The non-transfer class was lower with a 61.2% F1 (46.7% precision, 89% recall). These results show that factual findings projected from a small set of examples to the entire corpus are predictive of case outcomes in the entire corpus.

6.3 Predicting Findings from Case Text

In our next experiment, we measured our ability to predict tags in the Findings section from the case text, i.e., the first 5 sections of cases. We selected 3 Factors: ICS-CownsTM (Identical or Confusingly Similar, Complainant owns TradeMark), NRLI-PrimaFacieEst (No Rights or Legitimate Interests, Prima Facie Established), and BadFaith-DisruptBiz (Bad Faith, DisruptBusiness). These were the most frequently occurring findings in our corpus of projected tags.

To create the training data for this experiment, we used the tag projection method described above. The projected sentence annotations were interpreted as labels for a binary document classification task. Each finding label is skewed towards the negative class. The ICS label distribution is 75.8% negative and 24.2% positive, NRLI is 80.3% negative and 19.7% positive, finally Bad Faith label distribution is 74.2% negative and 25.8% positive.

The prediction accuracy is similar for all labels. Over 10 trials, Bad Faith and NRLI had an average AUC of 63.5% and 62.9% respectively, while ICS was a bit lower with 60.1% AUC. Examining the per class precision, recall and F1 metrics, Bad Faith had the strongest result with 79.8% F1 for the negative class (81.5% precision, 78.2% recall) and 46.3% F1 for the positive class (43.9% precision, 48.9% recall). ICS and NRLI had similar results. ICS had 79.5% F1, for the negative class (80.9% precision, 78.2% recall) and 40.2% F1 for the positive class (38.3% precision, 42.3% recall). Results for NRLI prediction are 78.1% for the negative class (86.5% precision, 71.1% recall) and 40.1% F1 for the positive class (31.7% precision, 54.6% recall). These results indicate that predicting individual findings from case text is more difficult that predicting the outcome of the case from the findings. However, even with out small initial annotated set we obtained highly promising preliminary results.

6.4 Summary

This section has described a methodology for inducing case features for decision prediction by exploiting regularities in case corpora. The use of these features for justification and explanation is beyond the scope of this paper. However, over 30 years of scholarship has been directed at techniques, including case-based reasoning (CBR), dialectical argumentation, and rule-based justification, for justification, explanation, and argumentation in cases represented in terms of outcome-relevant features. The SCALE approach is agnostic as to which of these techniques should be applied. The key research contribution of this work is enabling these techniques to apply to cases represented as text without requiring manual feature processing.

7 RELATED WORK

Several previous projects have addressed SCALE's goal of extracting factors from case texts for use in prediction or for other purposes. Ashley et al. 2009 [6] extracted CATO factors [5] from *squibs*, a form of case summary, using text classification [15] then predicted case outcomes by applying a machine learning algorithm, *Issue-Based Prediction* [16], to cases represented using those factors. SCALE differs from this approach in two important respects. First, SCALE uses features that arise from a linguist's annotations of the Decision portions of a representative sample of cases, whereas CATO factors were engineered and refined by multiple experts over a period of many years. Moreover, SCALE extracts factors from unrestricted free text from cases, unlike [6], which processed only squibs, which were themselves produced by domain experts. Accordingly, SCALE has both much less onerous feature development and broader applicability. Features defined and extracted by SCALE's much more highly automated approach are almost certain to be less precise than CATO factors. However, SCALE makes system development for new domains tractable in a manner that previous, more labor-intensive

approaches are not. Other approaches to automated features extraction from free text include efforts for automated annotation described in [34], event detection for indexing as described in [22], and identifying cited facts and principles in legal text [31]. McCarty (2018) [23] advocated using word embeddings to augment hand-coded extraction patterns; SCALE instead uses example texts from Findings and Decision sections.

The argumentation-mining community has produced a number of efforts addressing the task of identifying argumentation units from free text (see generally [21]). However, despite their central role in legal argumentation and justification, case features, including factors, are not argument units *per se*. Accordingly these techniques are not directly applicable to the task of identifying relevant case features.

8 CONCLUSIONS AND FUTURE WORK

Computational techniques for explainable legal problem solving have existed for many years, but broad adoption of these techniques has been impeded by their requirement for manual case representation. The rise of large-scale text analytics and machine learning promised a way to finesse this obstacle, but the limited explanatory capability to these approaches has limited their adoption in law where, for deep institutional reasons, decisions must be justified in terms of authoritative legal sources.

This paper has described two approaches to exploiting the scalability of machine learning while retaining explanatory capability. An evaluation of the first approach, NFE, failed to establish significant improvement in decision accuracy from decision support in the form of highlighting of case text based on attention weights. This experiment doesn't conclusively establish that this approach can not be made to work, but it suggests the limits of explanation and justification based purely on text without explicit reference to legally-relevant concepts. This negative result motivated a second approach, SCALE (Semi-supervised Case Annotation for Legal Explanations), that exploits structural and semantic regularities in case corpora to identify textual patterns corresponding to annotations in a subset of cases. This approach obviates manual representation of any but a representative set of cases while at the same time enabling all of the more traditional techniques for explanation and justification.

A more complete evaluation of the SCALE approach on the WIPO data set will be possible when we have expanded our initial set of annotated case justification examples. In addition, our current projection technique does not consistently distinguish between positive and negative polarity instances. While vector-space similarity captures some forms of polarity expression, our current approach often conflates positive and negative findings. Simply identifying the relevant features of a case, e.g., the matters that the panel will be ruling on, can be valuable from the standpoint of decision support. However, accuracy in predictive, as opposed to issue identification, depends on accurately identifying the polarity of each finding. We are currently experimenting with techniques for assigning polarity to projected labels.

Our evaluation of SCALE used the statement of facts and contentions in each case as a proxy for actual case records, which may be both more complete and less orderly than the panel's own summary. We anticipate applying the SCALE methodology to an administrative agency starting in the next few months, which will provide a more realistic evaluation of its ability to scale to actual agency decision making.

This work was undertaken in the hope of developing a methodology equally applicable to routine administrative adjudications in agencies throughout the world. We hope that it will provide a model for how to apply contemporary computational linguistics techniques to homogeneous case corpora to produce explainable decision prediction systems with a minimum of manual case annotation.

ACKNOWLEDGMENTS

The MITRE Corporation is a not-for-profit company, chartered in the public interest. This document is approved for Public Release; Distribution Unlimited. Case Number 18-4558. ©2019 The MITRE Corporation. All rights reserved.

REFERENCES

[1] Al-Abdulkarim, L., Atkinson, K., Bench-Capon, T.J.M., Whittle, S., Williams, R., Wolfenden, C.: Noise induced hearing loss: An application of the angelic methodology. In: Legal Knowledge and Information Systems - JURIX 2017: The Thirtieth Annual Conference, Luxembourg, 13-15 December 2017. pp. 79–88 (2017)

[2] Alarie, B., Niblett, A., Yoon, A.: Using Machine Learning to Predict Outcomes in Tax Law (December 15 2017), available at SSRN: https://ssrn.com/abstract=2855977 or http://dx.doi.org/10.2139/ssrn.2855977

[3] Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. PeerJ CompSci (October 24 2016), https://peerj.com/articles/cs-93/

[4] Aleven, V., Ashley, K.: Evaluating a learning environment for case-based argumentation skills. In: Proceedings of the Sixth International Conference on Artificial Intelligence and Law. pp. 170–179. ACM Press, University of Melbourne, Melbourne, Australia (June 30–July 3, 1997)

[5] Aleven, V.A.W.M.M.: Teaching Case-based Argumentation Through a Model and Examples. Ph.D. thesis, University of Pittsburgh, Pittsburgh, PA, USA (1997), aAI9821228

[6] Ashley, K.D., Brüninghaus, S.: Automatically classifying case texts and predicting outcomes. Artif. Intell. Law 17(2), 125–165 (Jun 2009)

[7] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

[8] Bench-Capon, T.J.M., Dunne, P.E.: Argumentation in artificial intelligence. Artif. Intell. 171(10-15), 619–641 (Jul 2007)

[9] Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational Linguistics 22(1), 39–71 (Mar 1996)

[10] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. CoRR abs/1607.04606 (2016), http://arxiv.org/abs/1607.04606

[11] Boles, D.B., Adair, L.P.: The multiple resources questionnaire (mrq). Proceedings of the Human Factors and Ergonomics Society Annual Meeting 45(25), 1790–1794 (2001)

[12] Branting, L.K.: Reasoning with Rules and Precedents: A Computational Model of Legal Analysis. Kluwer Academic Publishers, Dordrect/Boston/London (2000)

[13] Branting, L.K., Yeh, A., Weiss, B., Merkhofer, E.M., Brown, B.: Inducing predictive models for decision support in administrative adjudication. In: AI Approaches to the Complexity of Legal Systems - AICOL International Workshops 2015-2017, Revised Selected Papers. Lecture Notes in Computer Science, vol. 10791, pp. 465–477. Springer (2017)

[14] Brooke, J.: SUS–a quick and dirty usability scale. Usability evaluation in industry 189(194), 4–7 (1996)

[15] Brüninghaus, S., Ashley, K.D.: Toward adding knowledge to learning algorithms for indexing legal cases. In: Proceedings of the 7th International Conference on Artificial Intelligence and Law. pp. 9–17. ICAIL '99, ACM, New York, NY, USA (1999), http://doi.acm.org/10.1145/323706.323709

[16] Bruninghaus, S., Ashley, K.D.: Predicting outcomes of case based legal arguments. In: Proceedings of the 9th International Conference on Artificial Intelligence and Law. pp. 233–242. ICAIL '03, ACM, New York, NY, USA (2003)

[17] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16, ACM, New York, NY, USA (2016)

[18] Gunning, D.: Defense advanced research projects agency (darpa) program information: Explainable artificial intelligence (xai) (2018), https://www.darpa.mil/program/explainable-artificial-intelligence, last visited December 26, 2018
[19] Hadfield, G.: Rules for a Flat World: Why Humans Invented Law and How to Reinvent It for a Complex Global Economy. Oxford University Press (2016)
[20] Katz, D.M., II, M.J.B., Blackman, J.: A general approach for predicting the behavior of the supreme court of the united states. PLoSOne 12(4) (2017)
[21] Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. ACM Trans. Internet Technol. 16(2), 10:1–10:25 (Mar 2016)
[22] Maxwell, K.T., Oberlander, J., Lavrenko, V.: Evaluation of semantic events for legal case retrieval. In: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval. pp. 39–41. ESAIR '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1506250.1506259
[23] McCarty, L.T.: Research Handbook on the Law of Artificial Intelligence, chap. Finding the right balance in artificial intelligence and law. Edward Elgar Publishing (2018)
[24] Mikolov, T., Yih, S.W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013). Association for Computational Linguistics (May 2013)
[25] Peterson, M., Waterman, D.: Rule-based models of legal expertise. In: Walters, C. (ed.) Computing Power and Legal Reasoning, pp. 627–659. West Publishing Company, Minneapolis, Minnesota (1985)
[26] Rissland, E., Skalak, D., Friedman, T.: Supporting legal arguments through heuristic retrieval. Artificial Intelligence and Law 4(1), 1–71 (1996)
[27] Rissland, E.L., Ashley, K.D., Branting, L.K.: Case-based reasoning and law. The Knowledge Engineering Review 20(3), 293–298 (2005)
[28] Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. CoRR abs/1509.00685 (2015), http://arxiv.org/abs/1509.00685
[29] Sauro, J., Dumas, J.S.: Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1599–1608. CHI '09 (2009)
[30] Sergot, M.J., Sadri, F., Kowalski, R.A., Kriwaczek, F., Hammond, P., Cory, H.T.: The British Nationality Act as a logic program. Commun. ACM 29(5), 370–386 (May 1986), http://doi.acm.org/10.1145/5689.5920
[31] Shulayeva, O., Siddharthan, A., Wyner, A.: Recognizing cited facts and principles in legal judgements. Artificial Intelligence and Law 25(1), 107–126 (Mar 2017)
[32] Sulea, O., Zampieri, M., Vela, M., van Genabith, J.: Predicting the law area and decisions of french supreme court cases. In: RANLP. pp. 716–722. INCOMA Ltd. (2017)
[33] Surdeanu, M., Nallapati, R., Gregory, G., Walker, J., Manning, C.: Risk analysis for intellectual property litigation. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Law. ACM, Pittsburgh, PA (June 6–10 2011)
[34] Wyner, A.Z., Peters, W.: Lexical semantics and expert legal knowledge towards the identification of legal case factors. In: JURIX. Frontiers in Artificial Intelligence and Applications, vol. 223, pp. 127–136. IOS Press (2010)
[35] The EXplainable AI in Law (XAILA) 2018 workshop. http://xaila.geist.re (December 12–14 2018), Groningen, The Netherlands
[36] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of NAACL-HLT. pp. 1480–1489 (2016)