# Scalable Methods for Annotating Legal-Decision Corpora

**Lisa Ferro, John Aberdeen, Karl Branting, Craig Pfeifer,**
**Alexander Yeh, Amartya Chakraborty**
The MITRE Corporation
{lferro, aberdeen, lbranting, cpfeifer, asy, achakraborty}@mitre.org

## Abstract

Recent research has demonstrated that judicial and administrative decisions can be predicted by machine-learning models trained on prior decisions. However, to have any practical application, these predictions must be explainable, which in turn requires modeling a rich set of features. Such approaches face a roadblock if the knowledge engineering required to create these features is not scalable. We present an approach to developing a feature-rich corpus of administrative rulings about domain name disputes, an approach which leverages a small amount of manual annotation and prototypical patterns present in the case documents to automatically extend feature labels to the entire corpus. To demonstrate the feasibility of this approach, we report results from systems trained on this dataset.

## 1 Introduction

Recent research has demonstrated that judicial and administrative decisions can be predicted by machine-learning models trained on prior decisions (Medvedeva et al., 2018). Predictive legal models have the potential to improve both the delivery of services to citizens and the efficiency of agency decision processes, e.g., by making benefits adjudications faster and more transparent, and by enabling decision-support tools for evaluating benefits claims.

The accuracy of predictive legal models is highest, and explanatory capability greatest, when the prior decisions are represented in terms of features manually engineered to express exactly the most relevant aspects of the prior case (Katz et al., 2017). However, this approach is not scalable. Alternatively, decisions can be predicted from case text alone, but these models typically lack explanatory capability (Aletras et al., 2016). Development of approaches for explaining decision predictions in terms of relevant case facts while minimizing manual feature engineering is critical for broad adoption of systems for legal case prediction.

Our approach to explainable legal prediction focuses on annotating the portions of written decisions that set forth the justification for the decision. We hypothesize that tag sets for these justifications can be used as features for explainable prediction. In a separate paper (Branting et al., 2019) we propose an approach that uses models trained on an annotated corpus to extract features that can be used for both outcome prediction and explanation in new cases. This paper focuses on development of the annotated corpus itself.

Our feature set makes use of two common elements in legal argumentation: *issues* and *factors*. In our usage, an "issue" is a formal element of a legal claim corresponding to a term, or "predicate," that occurs in an authoritative legal source, such as a statute, regulation, or policy, and that is cited in the decision portion of cases. For example, in jurisdictions in which the term "intoxication" occurs in a statute forbidding driving under the influence of alcohol or drugs (DUI), the predicate "intoxication" is an issue, and legal liability depends on whether this predicate is established at trial. "Slurred speech," by contrast, is a "factor" if the decision portion of one or more cases contains findings about "slurred speech" that justify conclusions about the issue of intoxication. This usage differs from Ashley (1991) and others in that our factors are not features developed by domain experts, but rather are classes of factual findings in case decisions denoted by common annotation tags. We surmise that these decision-

*Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 12–20
Minneapolis, Minesotta, June 7, 2019. ©2019 Association for Computational Linguistics

derived factors are amenable both to HYPO/CATO-like argumentation (Ashley, 1991 and Aleven, 1997) and to alternative machine-learning and inferential techniques.

This paper describes several approaches to lightweight and expedited corpus creation in support of explainable legal decision prediction. The methods are tested on formal written decisions about domain name disputes which are published by the World Intellectual Property Organization (WIPO). The first approach involves human annotators applying a three-layer schema for labeling argument elements, issues, and factors in the panel's findings and decisions on the case. This method is applied to a very small corpus of 25 documents. The second approach is applied to the entire corpus of over 16000 documents and employs a combination of automated preprocessing and human annotation for labeling the outcome for three principal issues in each WIPO case. A third layer of annotation is added by automatically projecting the argument element, issue, and factor annotations onto each sentence in each document of the entire corpus. We are making this a richly annotated corpus available to the research community via MITRE's GitHub space (https://github.com/mitre).

### 1.1 Prior work

Most early research in automated legal reasoning involved logical representations of legal rules (McCarty 2018). These systems often could justify conclusions in terms of rules, but two factors limited their adoption: (a) the challenges of accurately and scalably formalizing legal rules in computational logic, and (b) the difficulty of matching abstract predicates in rules (e.g., "nuisance") to case facts (e.g., "barking dog").

Research in legal Case-Based Reasoning (CBR) addressed these challenges by reasoning about the similarities and differences between the facts of a given new case and prior cases ("precedents"). The most influential approach to legal CBR involved factor-based argumentation (Ashley, 1991). For example, the CATO system (Aleven, 1997) employed a hierarchy of 26 factors organized into five higher level abstract concepts, or issues. Recent use of the CATO corpus to support automated identification of factors includes Wyner & Peters (2012) and Wyner (2010), who use GATE Teamware to perform manual annotation of factors. Al-Abdulkarim et al. (2015) annotate both factors and issues in the CATO corpus.

More recently, Sulea et al. (2017) attempted to automatically identify facts in the case description within the top 20 highest ranking bigrams and trigrams as defined by a classification model. While these spans of text were predictive of the area of law, they did not correspond to the facts of the case.

Our objective is a methodology that permits rapid development of explainable predictive systems in new domains. Accordingly, our case features are derived from the justification portion of texts of representative decisions—a readily accessible resource—rather than from the comparatively scarce resource of combined AI and legal expertise. We hypothesize that machine-learning models for deriving these features from the texts of new cases will permit explainable prediction, including both CATO-style factor analysis and other analytical techniques.

## 2 Data

Disputes over WWW domain name ownership are administrated by the United Nations' World International Property Organization (WIPO), under the Uniform Domain Name Dispute Resolution Policy (UDRP).[1] If a domain name has been previously registered by Party A, and Party B feels that the domain name rightly belongs to them instead, then Party B may file a complaint with WIPO, requesting that the domain name be transferred to them. Party A, the respondent, has the opportunity to respond to the complaint filed by Party B, the complainant. An independent panel of one or more individuals is assigned to review the case and make a ruling. The ruling is published on the WIPO website.[2]

The panel's written decision is divided into multiple sections, including the naming of the parties involved, the domain name(s) in dispute, a summary of the factual background, a summary of the complainant's and respondent's contentions, the panel's discussion of the foregoing information and the panel's legal findings based thereon, and

[1] See https://www.icann.org/resources/pages/policy-2012-02-25-en

[2] For example, see , e.g., https://www.wipo.int/amc/en/domains/search/text.jsp?case=D2016-1709

finally, the panel's decision on the case overall. Because of the fairly formulaic nature of these documents, they provide a rich source of data for developing automated algorithms that process information about legal issues, factors, and findings. The decision documents are freely downloadable from the WIPO website, with decisions dating back to 2007.

## 3 Annotation of Argument Elements, Issues and Factors

A key goal of our research is developing a repeatable 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, our approach requires only the linguistic skills necessary to annotate the decision portion of a representative subset of cases, a much more limited process.

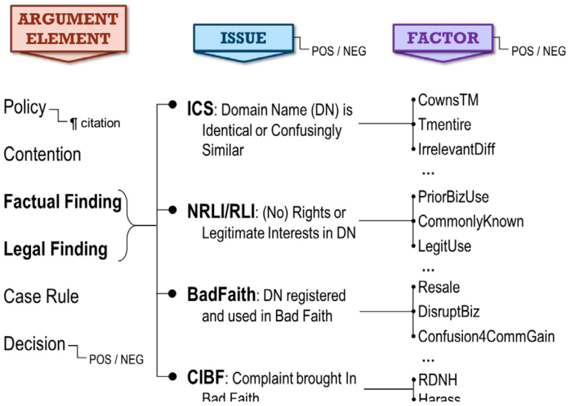


Figure 1: Annotation Scheme for WIPO Decisions

The annotation schema consists of three layers, shown in Figure 1: *Argument Elements*, *Issues*, and *Factors* (sub-issues). The top-layer, Argument Element, consists of six types: Policy, Contention, Factual Finding, Legal Finding, Case Rule, and the Decision on the case as a whole. 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. Each of these argument elements is generally found in all legal and administrative rulings, so by using these as the anchoring elements of the analysis scheme, our intent is that this approach will have utility in other domains.

We hypothesize that Factual Findings and Legal Findings will have the most predictive and explanatory power. Therefore, Factual Findings and Legal Findings are further categorized according to the *Issue* the panel is addressing. Contentions and Case Rules can also be labelled according to the issue they address.

The Issue tags include the three required elements that the complainant must establish in order to prevail in a WIPO case. These issues are documented in the Uniform Domain Name Dispute Resolution Policy, paragraph 4,[3] and 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 with respect to 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 no RLI (i.e., NRLI), it is very often the case that the panel considers evidence in support of the rights or legitimate interests of 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).[4]

Each of these issues can be further sub-categorized according to *Factors*, a sampling of which is shown in Figure 1. Factors are the elements which we hypothesize will prove most useful for explainable legal prediction. For example, whether or not the complainant owns rights in the trademark (CownsTM) is a critical factor in establishing the outcome of the first issue about the confusability of the domain name and trademark. In our annotation scheme, there are eight factors for the ICS issue, four factors for

[3] https://www.icann.org/resources/pages/policy-2012-02-25-en#4

[4] see 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 |
|---|---|---|
| D2012-1430 | 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 Panel thus finds that the disputed domain names are confusingly similar to the Complainant's registered trademarks | LEGAL_FINDING-ICS |
| D2012-1430 | Additionally, as several of the disputed domain names are used to host online shopping websites offering products similar to those of the Complainant, from which the Respondent presumably generates revenue, | FACTUAL_FINDING-NRLI-LegitUse subissue-polarity=negative |
| 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 | …the Respondent is clearly attempting to divert Internet traffic intended for the Complainant's website to its own for commercial gain by creating a likelihood of confusion as to the source or sponsorship of the Respondent's websites and products. | FACTUAL_FINDING-BadFaith-Confusion4CommGain |
| D2012-1430 | Such use constitutes bad faith under paragraph 4(b)(iv) of the Policy. | LEGAL_FINDING-BadFaith-Confusion4CommGain |
| 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 | Taking into account all of the above the Panel has no hesitation in finding that the present case amounts to RDNH by the Complainant. | LEGAL_FINDING-CIBF-RDNH |

Table 1: Example Annotations from WIPO Decisions

NRLI/RLI, and seven for BadFaith. Some of the factors are derived from the WIPO policy and some were discovered in a pilot annotation phase. 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 (not shown in Figure 1) to accommodate semantics that are not covered by the predefined tags, and there is a free-form Comment field which the annotator can use to capture ad hoc labels and enter notes.

A Citation attribute is used to capture the paragraph citation of Policy references, when they are explicitly made in the text. We plan to explore the citations as predictive features in future research. Finally, a Polarity attribute is used to capture positive/negative values for Decisions, Issues, and Factors.

Our three-layered annotation approach of labeling Argument Elements, Issues, and Factors relies on having clear divisions between the facts, the decision, and the decision justification. Aletras et al. (2016) found that the facts section of the case text provided the best predictors of case decision. It is our hypothesis that this methodology will be particularly useful in domains with a specific set of issues and justifications, for example, granting government benefits, or tenant/landlord disputes.

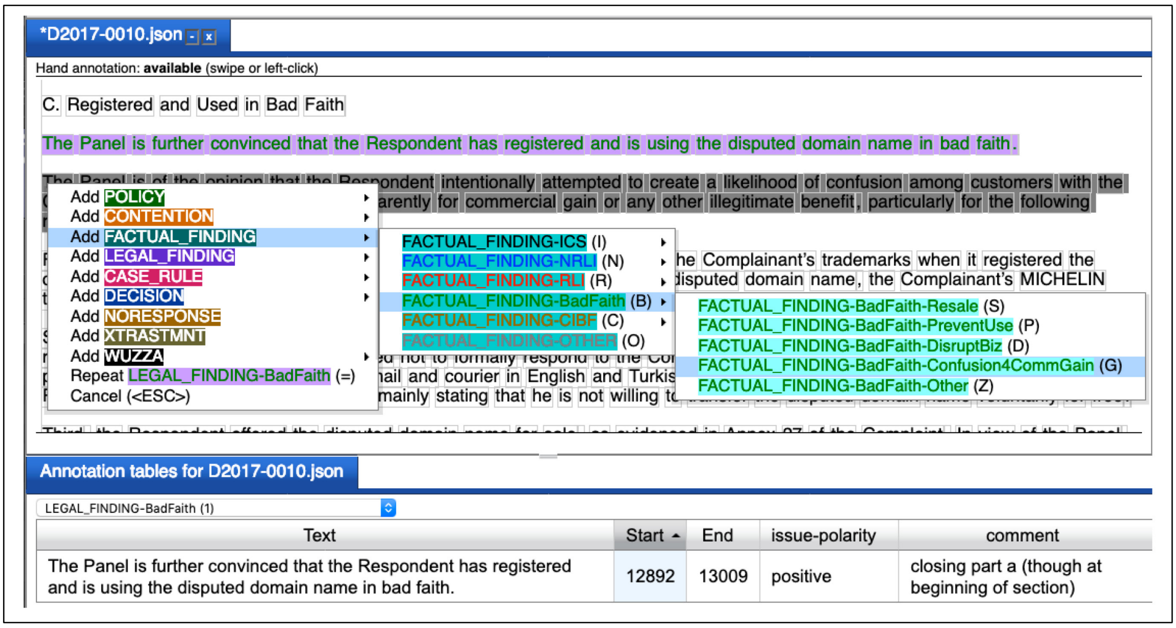
Table 1 shows eight typical annotations. Tags are preferentially applied to clauses and sentences, as opposed to shorter units such as noun phrases, in order to identify the complete linguistic proposition corresponding to the annotation label.

The MITRE Annotation Toolkit (MAT) is used to perform the annotation.[5] A screenshot is shown in Figure 2, illustrating the cascading menus that give the annotator quick access to the entire tag hierarchy.

### 3.1 Inter-Annotator Agreement

The manual annotation was performed by two individuals, who, while experienced in the creation of annotated corpora for developing natural language processing systems, have no formal training in the legal domain. Before performing the double annotation used for agreement measures, annotation guidelines were written and a set of six practice documents was identified, three with a positive outcome (the domain name was transferred to the complainant) and three with a

[5] Available for download at http://mat-annotation.sourceforge.net/



*D2017-0010.json
Hand annotation: **available** (swipe or left-click)
Add POLICY
Add CONTENTION
Add FACTUAL_FINDING
Add LEGAL_FINDING
Add CASE_RULE
Add DECISION
Add NORESPONSE
Add XTRASTMNT
Add WUZZA
Repeat LEGAL_FINDING-BadFaith (=)
Cancel (<ESC>)
FACTUAL_FINDING-ICS (I)
FACTUAL_FINDING-NRLI (N)
FACTUAL_FINDING-RLI (R)
FACTUAL_FINDING-BadFaith (B)
FACTUAL_FINDING-CIBF (C)
FACTUAL_FINDING-OTHER (O)
FACTUAL_FINDING-BadFaith-Resale (S)
FACTUAL_FINDING-BadFaith-PreventUse (P)
FACTUAL_FINDING-BadFaith-DisruptBiz (D)
FACTUAL_FINDING-BadFaith-Confusion4CommGain (G)
FACTUAL_FINDING-BadFaith-Other (Z)
Annotation tables for D2017-0010.json
LEGAL_FINDING-BadFaith (1)

C. Registered and Used in Bad Faith

The Panel is further convinced that the Respondent has registered and is using the disputed domain name in bad faith.

| Text | Start | End | issue-polarity | comment |
|---|---|---|---|---|
| The Panel is further convinced that the Respondent has registered and is using the disputed domain name in bad faith. | 12892 | 13009 | positive | closing part a (though at beginning of section) |

Figure 2: MITRE Annotation Tool (MAT)

negative outcome. These were doubly annotated in three trial phases, with the annotators meeting after each pair of documents to discuss differences and clarify the guidelines. Once the practice phase was complete, additional double-annotation was performed on a different set of six documents, which yielded 232 annotations. Agreement was measured on these annotations. Table 2 presents the inter-annotator agreement results, reported as percent agreement using Cohen's Kappa calculation (Cohen, 1960) vs. the raw agreement, shown in parentheses. As there were only two annotators, we do not compute inter-annotator agreement comparisons in a pair-wise fashion, i.e., for each annotator separately.

| | Argument Element | Issue | Factor | Factor Normalized for Other vs. Nil |
|---|---|---|---|---|
| **All Argument Element Types** | 75% (80%) | 80% (85%) | 68% (76%) | 74% (84%) |
| **Only Legal Findings & Factual Findings** | 57% (78%) | 80% (86%) | 69% (74%) | 75% (83%) |

Table 2: Inter-Annotator Agreement (Kappa vs. Raw)

Overall, the agreement was 75% on argument elements, 80% on issues, and 68% on factors. For Legal Findings and Factual Findings – features which we hypothesize will have greater predictive and explainable power – the levels of agreement on this set of six documents was lower for argument elements, at 57%, and did not differ significantly for issues and factors. We observed that one difference that lowers agreement occurs when one annotator chooses to specify "Other" as a factor label and the other annotator opts to not set a factor label at all (an alternative that is allowed by the guidelines). It is quite subjective whether the semantics of the clause warrant an "Other" factor label or no label. If we normalize the difference, allowing Other and Nil to be equivalant, the agreement on factors increases from 68% to 74% on all argument element tags and from 69% to 75% on Legal Findings and Factual Findings.

The WIPO administrative decisions exhibit a fair amount of variability in terms of clarity when it comes to assigning argument elements, issues, and factor labels. For example, on some subsets of files, the two annotators were able to achieve raw agreement as high as 99% on the Issue labels for Legal Findings and Factual Findings. We found that for the majority of cases that were doubly annotated, agreement was higher on the Legal Findings and Factual Findings than across all argument element types, a fact that is not reflected in Table 2, which contains totals for all documents that were doubly annotated. Thus, although some cases are more challenging to annotate than others, overall the quantitative results indicate that the task is tractable for non-legal experts.

### 3.2 Predicting Decisions from Mapped Tags

From the small set of 25 annotated documents (0.14% of the entire corpus), we are able to project the annotations to similar sentences throughout the entire corpus of documents.

This projection is accomplished through the use of word and sentence embeddings to find text that is semantically similar to the annotated text. 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 performed by determining the predictive accuracy of the mapped tags. 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.

The projection method is as follows. Word embeddings are trained on the tokenized corpus using FastText (Mikolov, 2018). FastText computes embeddings for character n-grams and then sums the character n-gram embeddings to compute the embedding for a word. The character embeddings are computed using a method similar to Word2Vec (Mikolov, 2013). FastText is beneficial for rare words, morphologically rich languages and smaller corpora. This yields one vector per token that captures the semantics of the word through the surrounding context.

The resulting word embeddings are then 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. 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. We attempted to compensate for this limitation by incorporating polarity annotations into the projected tags. The annotation convention was the polarity attribute was assumed to be "positive" if not explicitly annotated. Out of 890 annotations, 173 (19.4%) contained negative polarity and 717 contained positive polarity. The polarity attribute was extracted for each annotation and incorporated into the projected tag. The sentences are then clustered into 512 clusters by their embeddings. The clusters establish neighborhoods of 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 (Chen and Guestrin, 2016), an efficient implementation of gradient boosted machines, is used in all prediction tasks in this work. These are preliminary results, and we continue to iterate to improve the outcomes.

As the outcome decision labels are highly skewed, 91% positive (14591 cases), 9% negative (1407 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 micro-averaged over the 10 trials.

In a separate paper (Branting et al., 2019) we report on several experiments that make use of the projected tags. As proof of concept for the methodology described in this paper, we report only on the results for predicting case outcomes. The results are preliminary, intended solely to demonstrate the feasibility of the approach.

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

78.5% and a standard deviation of 0.01. The positive class, the majority class in this dataset, earned a 90% F1 (97% precision and 83.9% recall). The negative class was lower with a 42.9% F1 (30.4% precision, 73.1% recall). This experiment indicates that tags mapped from a modest set of annotated cases are sufficient to express the decisions in the Findings section.

## 4 Annotation of legal rulings on issue outcome in WIPO Decisions

In a WIPO case, the complainant needs to prevail in each of the three primary issues – ICS, NRLI, and Bad Faith – in order for there to be a positive outcome, i.e., the domain name is transferred to the complainant. Being able to accurately identify the outcome of each issue is therefore useful in predicting the outcome of the case overall. Issue-level outcomes take the form of legal findings (e.g., see the second row in Table 1), and are typically found within easily identifiable sub-sections of the panel's decision. They often appear as the last sentence in the issue-level sub-sections – a pattern we were able to exploit, but only to a limited degree, as described below. The manner in which the legal finding is stated varies as well. As a result, a fully automated approach to issue-level outcome annotation was not possible, and manually annotating the corpus would be prohibitively time consuming for the entire set of over 48000 issue outcomes (16000 cases x 3). We therefore used a multi-step interactive approach to annotate the issue outcomes, described next.

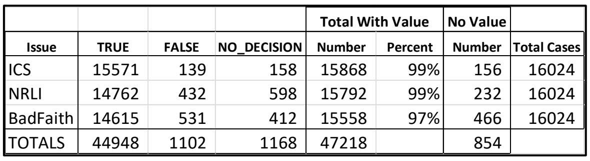
Approximately 90% of the cases have a positive outcome, so the first step was to automatically annotate positive cases with a positive outcome for each of the three issues. This left 10% of the cases that have a negative overall outcome, and which could potentially have negative outcomes in one, two, or all three of the issues – up to approximately 5000 issue outcomes in total. In those cases where the sub-section on an issue could not be automatically located in the panel's decision, it was temporally labelled as having a missing value for the issue outcome. A few hundred Bad Faith issue outcomes were manually annotated before it was deemed to take too long.

Next, we extracted all the unique last sentences in the issue-level subsections still missing an outcome annotation. This gave us approximately 3000 sentences which we manually annotated with one of the following values:

- True (positive outcome for the issue)
- False (negative outcome for the issue)
- No_Decision (The panel asserts that it does not need to make a decision on this issue because some other issue has a negative outcome.)
- ? (does not describe an outcome)

The manual annotation revealed some discrepancies which needed to be corrected, for example, cases with an overall positive outcome that had issues with a negative outcome and issue outcomes appearing outside their designated subsection.

Table 3 summarizes the current state of the issue-level annotation for the WIPO corpus of 16024 cases.



| | | | | Total With Value | | No Value | |
|---|---|---|---|---|---|---|---|
| Issue | TRUE | FALSE | NO_DECISION | Number | Percent | Number | Total Cases |
| ICS | 15571 | 139 | 158 | 15868 | 99% | 156 | 16024 |
| NRLI | 14762 | 432 | 598 | 15792 | 99% | 232 | 16024 |
| BadFaith | 14615 | 531 | 412 | 15558 | 97% | 466 | 16024 |
| TOTALS | 44948 | 1102 | 1168 | 47218 | | 854 | |

Table 3: Issue-Level Outcome Annotations

Between 97% and 99% of the corpus has been annotated for issue-level outcomes, depending on the issue, with 845 issue outcomes yet to be resolved.

For those cases that could not be annotated as True, False, or No_Decision, we are currently in the process of analyzing additional patterns that can be exploited automatically. We are also performing additional automated quality control checks that look for inconsistencies, e.g., the overall case outcome being false, but none of the issues are false.

In this section we have described a methodology for annotating a large corpus for issue-level decisions. While the exact approach is necessarily dependent on specifics that are unique to the WIPO domain, our expectation is that it is generalizable to other datasets. For instance, not all legal and administrative rulings have clearly identifiable case-level decisions, but when they do, they can be automatically extracted and then used to infer issue-level decision values. Other prototypical patterns can be used to automate the annotation, and the more complicated texts can be reserved for human review.

### 4.1 Issue-Outcome Prediction

We have begun experimenting with the issue-level outcome annotation, testing different machine-learning approaches for predicting issue outcomes based on these annotated sentences. The task is to predict whether a given sentence favors the respondent (a negative outcome), favors the complainant (a positive outcome), or states that the panel is not making a decision for this issue. A subset of the corpus was divided into 1438 training samples and 709 test samples. The best performance achieved thus far has been from an approach utilizing a 300-dimensional Word Embedding from FastText, with 1 million word vectors trained on Wikipedia in 2017 (Mikolov, 2017). Results are shown in Table 4.

| | Precision | Recall | F1 |
|---|---|---|---|
| **Negative Outcome** | 0.89 | 0.95 | 0.92 |
| **Positive Outcome** | 0.96 | 0.93 | 0.94 |
| **No_Decision** | 0.98 | 0.93 | 0.95 |

Table 4: Issue-Level Prediction Scores

## 5 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 feature labeling. The rise of large-scale text analytics and machine learning promised a way to finesse this obstacle, but the limited explanatory capability of these approaches has limited their adoption in law where decisions must be justified in terms of authoritative legal sources.

This paper has described three approaches to exploiting minimalist knowledge engineering in the form of an extremely small corpus annotated by non-legal experts and using that annotation and regularities discernible in the data to automatically augment and extend these annotations. We have provided proof-of-concept of the quality and utility of these annotations by reporting preliminary results on issue-level and case-level decision prediction algorithms.

We anticipate that the accuracy of annotation projection can be improved by use of improved embeddings methods. Since this work has started, very large pre-trained language models have been released, including ELMO (Peters, 2018), BERT (Devlin, 2018) and GPT-2 (Radford, 2019). Future work should use these pre-trained models to create embeddings for sentences and annotations to improve the tag projection. These embedding algorithms capture greater nuance of language than FastText/Word2Vec (possibly including word sense), and their pre-trained models are built from massive data collections processed on massive compute clusters.

In future work, the domains of particular interest to us include disability benefit claims, immigration petitions, landlord-tenant disputes, and attorney misconduct complaints. The high volumes of these types of cases mean that large training sets are available and that agencies have an incentive to consider technologies to improve decision processes. We anticipate applying the annotation methodologies described in this paper to an administrative agency starting in the next few months, which will provide a more realistic evaluation of its ability to support system development in the service of actual agency decision making.

## Acknowledgments

We would like to thank the anonymous reviewers for their feedback and suggestions.

## References

Al-Abdulkarim, Latifa, Katie Atkinson, and Trevor Bench-Capon. 2015. Factors, Issues and Values: Revisiting Reasoning with Cases. In *Proceedings of the 15th International Conference on Artificial intelligence and Law* (pp. 3-12). ACM.

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. *PeerJ Computer Science* 2: e93. https://peerj.com/articles/cs-93/ .

Aleven, Vincent. 1997. *Teaching Case-Based Argumentation Through a Model and Examples*. Ph.D. thesis, University of Pittsburgh.

Aleven, Vincent. 2003. Using Background Knowledge in Case-based Legal Reasoning: a Computational Model and an Intelligent Learning Environment. *Artificial Intelligence* 150.1-2: 183-237.

Ashley, Kevin D. 1991. *Modeling Legal Arguments: Reasoning with Cases and Hypotheticals*. MIT Press.

Branting, Karl, Craig Pfeifer, Lisa Ferro, Alex Yeh, Brandy Weiss, Mark Pfaff, Amartya Chakraborty, and Bradford Brown. 2019. Semi-supervised Methods for Explainable Legal Prediction. To appear, *Proceedings of the 19th International*

*Conference on AI and Law (ICAIL 2019)*, Montreal, Canada, June 17-21, 2019.

Chen, Tianqi, and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794. http://arxiv.org/abs/1603.02754.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pp. 37-47.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR abs/1810.04805 (2018)*.

Katz, Daniel Martin, Michael J. Bommarito II, and Josh Blackman. 2017. A General Approach for Predicting the Behavior of the Supreme Court of the United States. *PloS One*, 12:4.

McCarty, L. Thorne. 2018. Finding the right balance in Artificial Intelligence and Law. In *Research Handbook on the Law of Artificial Intelligence*, Edward Elgar Publishing.

Medvedeva, Masha, Michel Vols, and Martijn Wieling. 2018. Judicial Decisions of the European Court of Human Rights: Looking into the Crystal Ball. In *Proceedings of the Conference on Empirical Legal Studies*.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in Pre-training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018),* pp. 52-55. http://aclweb.org/anthology/L18-1008.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111-3119.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke S. Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of NAACL-HLT*, pp. 2227-2237. https://aclweb.org/anthology/N18-1202

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Sulea, Octavia-Maria, Marcos Zampieri, Mihaela Vela and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (*RANLP 2017)*, pp. 716-722. https://aclweb.org/anthology/papers/R/R17/R17-1092/

Wyner, Adam, and Wim Peters. 2012. Semantic Annotations for Legal Text Processing using GATE Teamware. In *Semantic Processing of Legal Texts* (SPLeT-2012) Workshop Programme, p. 34.

Wyner, Adam Z. 2010. Towards Annotating and Extracting Textual Legal Case Elements. In *Informatica e Diritto: Special Issue on Legal Ontologies and Artificial Intelligent Techniques* 19.1-2: 9-18.