Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning

Vern R. Walker
Director, Research Laboratory for Law, Logic & Technology (LLT Lab)
Maurice A. Deane School of Law, Hofstra University
Hempstead, New York, USA
vern.r.walker@hofstra.edu

Krishnan Pillaipakkamnatt
Chair, Department of Computer Science
Fred DeMatteis School of Engineering and Applied Science, Hofstra University
Hempstead, New York, USA
krishnan.pillaipakkamnatt@hofstra.edu

Alexandra M. Davidson
Research Laboratory for Law, Logic & Technology (LLT Lab)
Maurice A. Deane School of Law, Hofstra University
Hempstead, New York, USA
lltlab@hofstra.edu

Marysa Linares
Research Laboratory for Law, Logic & Technology (LLT Lab)
Maurice A. Deane School of Law, Hofstra University
Hempstead, New York, USA
lltlab@hofstra.edu

Domenick J. Pesce
Research Laboratory for Law, Logic & Technology (LLT Lab)
Maurice A. Deane School of Law, Hofstra University
Hempstead, New York, USA
lltlab@hofstra.edu

ABSTRACT

Automatically mining patterns of reasoning from evidence-intensive legal decisions can make legal services more efficient, and it can increase the public's access to justice, through a range of use cases (including semantic viewers, semantic search, decision summarizers, argument recommenders, and reasoning monitors). Important to these use cases is the task of automatically classifying those sentences that state whether the conditions of applicable legal rules have been satisfied or not in a particular legal case. However, insufficient quantities of gold-standard semantic data, and the high cost of generating such data, threaten to undermine the development of such automatic classifiers. This paper tests two hypotheses: whether distinctive phrasing enables the development of automatic classifiers on the basis of a small sample of labeled decisions, with adequate results for some important use cases, and whether semantic attribution theory provides a general methodology for developing such classifiers. The paper reports promising results from using a qualitative methodology to analyze a small sample of classified sentences (N = 530) to develop rule-based scripts that can classify sentences that state findings of fact ("Finding Sentences"). We compare those results with the performance of standard machine learning (ML) algorithms trained and tested on a larger dataset (about 5,800 labeled sentences), which is still relatively small by ML standards. This methodology and these test results suggest that some access-to-justice use cases can be adequately addressed at much lower cost than previously believed. The datasets, the protocols used to define sentence types, the scripts and ML codes will be publicly available.

ACM Reference format:

Vern R. Walker, Krishnan Pillaipakkamnatt, Alexandra M. Davidson, Marysa Linares and Domenick J. Pesce. 2019. Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019)*, Montreal, QC, Canada, 10 pages.

1 Introduction

Automated argument mining would add greatly to productivity in legal practice, and it could increase access to justice in many legal areas through a range of use cases. As a first use case, a web-based semantic viewer might automatically highlight for the user those sentences and patterns that are of interest in argumentation. For example, a semantic viewer might open a new decision document and provide filters that the user could select to highlight only the conclusions, or the evidence, or the stated reasoning from evidence to conclusions.

Second, we could create semantic search tools, which would use components of reasoning to search through hundreds of thousands of decisions in plain-text format, retrieve those decisions similar to a new case (e.g., those with similar issues to be proved, and similar types of evidence available for proving them), rank the similar decisions in order of greatest similarity, and then display the portions to read (using a semantic viewer).

Third, the capability of extracting from published decisions the major conclusions, the intermediate reasoning, and the evidentiary basis for a decision would also provide the components of an informative summary of that decision. A decision summarizer

In: Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019), June 21, 2019, Montreal, QC, Canada.

Published at http://ceur-ws.org

could analyze a decision and create a digest or summary for the human reader.

Fourth, an automatic argument miner could extract successful and unsuccessful patterns of reasoning from thousands of past decisions, and then it could generate suggestions for new arguments in new cases. Such an argument recommender could assist attorneys, non-attorneys, and judges in processing new cases.

Fifth, such an automated argument miner could monitor cases as they are being litigated, compare evolving arguments with patterns of reasoning that have been successful or unsuccessful in the past, and detect possible outliers (arguments likely to fail or decisions likely to be incorrect). A reasoning monitor could also maintain statistics and trends for patterns of reasoning, and it could predict probabilities of success for new cases.

Provided the data derived from argument mining is valid and predictive of real-case outcomes, such tools could assist alternative dispute resolution and increase efficiency within the legal system. Such automated, evidence-based tools (semantic viewers, semantic search, decision summarizers, argument recommenders, and reasoning monitors) could also assist non-lawyers when they represent themselves in cases where a lawyer is not available.

Producing such a range of tools, however, faces several challenges. One challenge is whether machine learning (ML) tools can be effective at automating such argument mining. First, there is the problem of the available quantity of gold-standard data for training and testing. Supervised ML may require such a large quantity of accurately labeled documents that there are not sufficient resources to generate it, in all areas of law where argument mining is desirable. While semi-supervised ML, using large quantities of unlabeled data and small quantities of labeled data, offers more promise, even that approach requires trained human classifiers. Especially in legal areas where outcomes may not economically support the hiring of a lawyer (e.g., veterans' disability claims or immigration asylum claims), there may be little financial incentive to create such a great quantity of ground-truth annotated data. Moreover, some areas of law (such as vaccine-injury compensation decisions) may not even produce a sufficient quantity of cases bearing on a particular issue, even if we were to annotate it all.

Moreover, there is the challenge of ensuring data validity. In order to create effective tools, especially tools that predict future outcomes given past decisions, the data upon which the tools are trained must accurately reflect what we believe it measures. But appropriately annotating the components of legal reasoning requires an adequate theory of legal reasoning, a sufficient number of trained annotators, and adequate quality assurance. Also, to inspire trust in ML outputs, the models must be transparent and understandable.

Finally, there is the challenge of developing and testing adequate classification type systems for legal arguments [34, 38]. Unsupervised ML has the challenge of producing useful clusters, especially if the components of legal argument are poorly understood, the components vary depending on the use case, and the components are different for different practitioners. How to classify argument types is therefore a problem that any effort to create gold-standard data must address.

This paper reports on preliminary research that addresses all of these problems. Our main hypothesis is that reports of fact-finding in legal adjudicatory decisions might employ such regular and distinctive phrasing that even rule-based scripts based on a very small sample, as well as ML models trained on larger samples, can perform adequately for many valuable use cases. Our second hypothesis is that attribution theory from linguistics can be extended to argument mining, as a method for creating semantic types and automatically identifying them in legal decisions. There is reason to think that such an approach will be transferable to adjudicatory decisions in many substantive areas of law.

Using an annotated dataset of U.S. decisions as the gold standard, we investigated a methodology for qualitatively studying a very small sub-sample of such decisions, developing rule-based scripts, and quantitatively testing the script performance. We compared those outcomes against the performance of standard supervised ML models trained on larger samples from the same dataset. Our study addresses data quantity by employing very small datasets; it addresses data validity by employing quality-assurance protocols and publishing those protocols and the resulting data; and it addresses annotation type systems by explaining their derivation. The paper reports promising results relative to important use cases, and it lays out a methodology that should be transferable to many areas of law at a relatively small cost – thus helping to improve access to justice. We make publicly available the annotated dataset, the quality-assurance protocols, the scripts and the ML settings, at https://github.com/LLTLab/VetClaims-JSON.

After we discuss prior related work in Section 2, we describe our dataset and how we used it in both our script and ML experiments (Section 3). Section 4 describes the qualitative-study experiments and their results, while Section 5 describes the ML experiments and their results. Section 6 contains our general discussion of these combined results and our future work.

2 Prior Related Work

The context for the work reported in this paper is the goal of automated argument mining from adjudicatory legal decisions. Such argument mining would automatically extract the evidence assessment and fact-finding reasoning found in adjudicatory decisions, for the purpose of identifying successful and unsuccessful units of evidentiary argument. Researchers generally identify an argument as containing a conclusion or claim, together with a set of one or more premises. [21, 40, 30, 15, 32]

A first level of analysis is to classify the rhetorical roles of sentences for argument mining – that is, assigning sentences roles as either premise or conclusion. Prior work on classifying such rhetorical roles in adjudicatory decisions includes: applying machine learning to annotate sentence types in vaccine-injury compensation decisions [3, 4, 5, 10]; assigning rhetorical roles to sentences in Indian court decisions [27]; classifying sentences as argumentative in the Araucaria corpus, including newspapers and court reports [19]; automatically summarizing legal judgments of the U.K.'s House of Lords, in part by classifying the rhetorical

status of sentences [13]; annotating sentences as containing legal principles or facts in common-law reports [29]; and using statistical parsing as the input for computing quasi-logical forms as deep semantic interpretations of sentences in U.S. appellate court decisions [17]. Al-Abdulkarim et al. provide one overview of statement types involved in legal reasoning in cases, from evidence to a verdict [1]. The approach we describe in this paper utilizes a type system of rhetorical roles developed to annotate any fact-finding decision, and we compare script and ML classifiers for the rhetorical roles of sentences. Moreover, it is not common for datasets to be publicly available, together with protocols for data generation, scripts and codes, to enable confirmation of data accuracy and replication of results.

A recent article compared two experiments in automated classification of legal norms from German statutes, with regard to their semantic type: (1) a rule-based approach using hand-crafted pattern definitions, and (2) an ML approach. [39] (For similar work on Dutch laws, see [16].) The performance metrics for the two experiments were comparable on a dataset of manually-labeled statements. While this study is highly relevant to our work, there are distinct differences. We develop a qualitative methodology for developing classification features of sentences in adjudicatory decisions (not statutes), according to their rhetorical role (not norm type), for the purpose of automated argument mining. Our methodology is general, and it should be transferable to adjudicatory decisions in any substantive area of law.

To identify the rhetorical roles of sentences, we employ an extension of the semantic theory of attribution analysis. Attribution, in the context of argument mining, is the descriptive task of determining which actor is asserting, assuming or relying upon which propositions, in the course of presenting reasoning or argument. Although attribution is a classic problem area in natural language processing generally [7, 14, 22, 23], there has been limited work on attribution in respect to argument mining from legal documents. Grover et al. reported on a project to annotate sentences in House of Lords judgments for their argumentative roles [11]. Two tasks were to attribute statements to the Law Lord speaking about the case or to someone else (attribution), and to classify sentences as formulating the law objectively vs. assessing the law as favoring a conclusion or not favoring it (comparison). This work extended the work of [31] on attribution in scientific articles. A broader discussion of attribution within the context of legal decisions is found in [34]. Unlike the adjudicatory decisions used in our study, the House of Lords judgments studied by [11] treated facts as already settled in the lower courts. Our study appears to be unique in using attribution analysis to help classify the rhetorical roles of sentences in the evidence assessment portions of adjudicatory texts.

We have also developed classification protocols (classification criteria and methods) for each rhetorical role. We use protocols to give precise meaning to the semantic type, to train new annotators, and to review the accuracy of human annotations. We also use such protocols to guide the development of the features or rule-based scripts for automatically classifying legal texts (e.g., [28]). Stab and Gurevych have classified such features into 5 groups [30]. For example, the main verb of a finding sentence tends to be in present tense, while the main verbs of evidence sentences tend to be in past tense. Features derived from the protocols can drive the application of high-precision / low-recall techniques of the kind used successfully by [15], which we argue is the performance desired for certain use cases but not others.

3 The Datasets

We developed a common dataset to use in comparing the classification performance of rule-based script classifiers with the performance of ML models. This section describes that dataset and how it was used.

3.1 The BVA PTSD Dataset

We analyzed 50 fact-finding decisions issued by the U.S. Board of Veterans' Appeals ("BVA") from 2013 through 2017 (the "PTSD dataset"). We arbitrarily selected those decisions from adjudicated disability claims by veterans for service-related post-traumatic stress disorder (PTSD). PTSD is a mental health problem that some people develop after experiencing or witnessing a traumatic event, such as combat or sexual assault. Individual claims for compensation for a disability usually originate at a Regional Office ("RO") of the U.S. Department of Veterans Affairs ("VA"), or at another local office across the country [2, 20]. If the claimant is dissatisfied with the decision of the RO, she may file an appeal to the BVA, which is an administrative appellate body that has the authority to decide the facts of each case based on the evidence [20]. The BVA must provide a written statement of the reasons or bases for its findings and conclusions, and that statement "must account for the evidence which [the BVA] finds to be persuasive or unpersuasive, analyze the credibility and probative value of all material evidence submitted by and on behalf of a claimant, and provide the reasons for its rejection of any such evidence." *Caluza v. Brown*, 7 Vet. App. 498, 506 (1995), *aff'd*, 78 F.3d 604 (Fed. Cir. 1996).

The veteran may appeal the BVA's decision to the U.S. Court of Appeals for Veterans Claims (the "Veterans Court") [20], but the standard of review for issues of fact is very deferential to the BVA. In order to set aside a finding of fact by the BVA, the Veterans Court must determine it to be "clearly erroneous." [20] And although either the claimant or the VA may appeal a Veterans Court decision to the U.S. Court of Appeals for the Federal Circuit, the Federal Circuit may only review questions of law, such as a constitutional challenge, or the interpretation of a statute or regulation relied upon by the Veterans Court. [2, 20] Except for constitutional issues, it "may not review any 'challenge to a factual determination' or any 'challenge to a law or regulation as applied to the facts of a particular case.'" *Kalin v. Nicholson*, 172 Fed.Appx. 1000, 1002 (Fed.Cir. 2006). Thus, the findings of fact made by the BVA are critical to the success or failure of a veteran's claim.

The BVA's workload has increased dramatically in the past decade, reaching 85,288 decisions in fiscal year 2018. [6, p. 32] The vast majority of appeals (96%) considered by the BVA involve claims for compensation. [6, p. 31] Therefore, identifying the

patterns of factual reasoning within the decisions of the BVA presents a significant challenge for automated argument mining.

For each of the 50 BVA decisions in our PTSD dataset, we extracted all sentences addressing the factual issues related to the claim for PTSD, or for a closely-related psychiatric disorder. This set of sentences ("PTSD-Sent") is the dataset on which we conducted our experiments. The "Reasons and Bases" section of the decision is the longest section, containing the Board's statement of the evidence, its evaluation of that evidence, and its findings of fact on the relevant legal issues.

3.1.1 Rhetorical Roles of Sentences in the PTSD-Sent Dataset

For the purpose of identifying reasoning or argument patterns, we focus primarily on sentences that play one of three rhetorical roles in evidence assessment: the **finding of fact**, which states whether a propositional condition of a legal rule is determined to be true, false or undecided; the **evidence** in the legal record on which the findings rest, such as the testimony of a lay witness or a medical record; and the **reasoning** from the evidence to the findings of fact. Identifying the sentences that have those roles within adjudicatory decisions, however, presents special problems. Such decisions have a wide diversity of roles for sentences: e.g., stating the legal rules, policies and principles applicable to the decision, as well as providing citations to authority; stating the procedural history of the case, and the rulings on procedural issues; summarizing the evidence presented and the arguments of the parties based on that evidence; and stating and explaining the tribunal's findings of fact based on that evidence. [37] Thus, BVA decisions pose the challenge of classifying rhetorically important types of sentence and distinguishing them from other types of sentence.

The following are the 5 rhetorical roles that we used to classify sentences in the PTSD-Sent dataset. Sentences were classified manually by teams of 2 trained law students, and they were curated by a law professor with expertise in legal reasoning. Data validity is open to scrutiny because the data will be publicly available.

Finding Sentence. A Finding Sentence is a sentence that primarily states a "finding of fact" – an authoritative conclusion of the trier of fact about whether a condition of a legal rule has been satisfied or not, given the evidence in the case. An example of a Finding Sentence is: "*The most probative evidence fails to link the Veteran's claimed acquired psychiatric disorder, including PTSD, to active service or to his service-connected residuals of frostbite.*" (BVA1340434)[1]

Evidence Sentence. An Evidence Sentence primarily states the content of the testimony of a witness, states the content of documents introduced into evidence, or describes other evidence. Evidence sentences provide part of the premises for findings of fact. An example of an Evidence Sentence is: "*The examiner who conducted the February 2008 VA mental disorders examination opined that the Veteran clearly had a preexisting psychiatric disability when he entered service.*" (BVA1303141)

Reasoning Sentence. A Reasoning Sentence primarily reports the trier of fact's reasoning underlying the findings of fact (therefore, a premise). Such reasoning often involves an assessment of the credibility and probative value of the evidence. An example of a Reasoning Sentence is: "*Also, the clinician's etiological opinions are credible based on their internal consistency and her duty to provide truthful opinions.*" (BVA1340434)

A unit of argument or reasoning within evidence assessment is usually composed of these three types of sentence (finding, evidence, and reasoning). The "Reasons and Bases" section of a BVA decision generally also includes two other types of sentence (those stating legal rules and citations), which must be distinguished from the first three. Unlike the case-specific elements of evidence, reasoning and findings, the legal rules and citations are often the same for tens of thousands of cases, even though the sentences stating those rules and citations can be highly variable linguistically, depending upon the writing style of the judge.

Legal-Rule Sentence. A Legal-Rule Sentence primarily states one or more legal rules in the abstract, without stating whether the conditions of the rule(s) are satisfied in the case being decided. An example of a Legal-Rule Sentence is: "*Establishing direct service connection generally requires medical or, in certain circumstances, lay evidence of (1) a current disability; (2) an in-service incurrence or aggravation of a disease or injury; and (3) a nexus between the claimed in-service disease or injury and the present disability.*" (BVA1340434)

Citation Sentence. A Citation Sentence references legal authorities or other materials, and usually contains standard notation that encodes useful information about the cited source. An example is: "*See Dalton v. Nicholson, 21 Vet. App. 23, 38 (2007); Caluza v. Brown, 7 Vet. App. 498, 511 (1995), aff'd per curiam, 78 F.3d 604 (Fed. Cir. 1996).*" (BVA1340434)

The frequencies of sentence rhetorical types within the PTSD-Sent dataset are shown in Table 1.

Rhetorical Type	Frequency
Finding Sentence	490
Evidence Sentence	2,419
Reasoning Sentence	710
Legal-Rule Sentence	938
Citation Sentence	1,118
Other Sentences	478
Total	6,153

Table 1. Frequency of Sentences in PTSD-Sent Dataset, by Rhetorical Type

For each rhetorical role, a protocol provides a detailed definition of the role, as well as methods and criteria for manually classifying sentences, and illustrative examples. Such protocols furnish materials not only for training annotators and for conducting quality assurance of data validity, but also for developing rule-based scripts that help automate the classification process. In this

[1] We cite decisions by their BVA citation number, e.g., "BVA1302544". Decisions are available from the VA website: https://www.index.va.gov/search/va/bva.jsp.

paper, we use initial caps in referring to a specific semantic type that is defined by a protocol (e.g., "Finding Sentence"), in contrast to a reference to a corresponding general concept (e.g., a finding sentence). The protocols for these five rhetorical roles will be made publicly available, along with the PTSD-Sent dataset.

3.1.2 "Finding Sentences" as Critical Connectors

"Finding Sentences" (as defined in Section 3.1.1 above) are critical connectors in argument mining. They connect the relevant evidence and related reasoning (which function as premises) to the appropriate legal issue, and they state whether a proponent's proof has been successful or not (the conclusion of the reasoning). Our experiments test the automatic classification of Finding Sentences, as distinct from the other sentence roles.

The governing substantive legal rules state the factual issues to be proved – that is, the conditions under which the BVA is required to order compensation, or the BVA is prohibited from ordering compensation. A legal rule can be represented as a set of propositions, one of which is the conclusion and the remaining propositions being the rule conditions [35, 18]. Each condition can in turn function as a conclusion, with its own conditions nested within it [37]. The resulting set of nested conditions has a tree structure – with the entire representation of the applicable legal rules being called a "rule tree" [35]. A rule tree integrates all the governing rules from statutes, regulations, and case law into a single, computable system of legal rules.

Figure 1 shows the highest levels of the rule tree for proving that a veteran's PTSD is "service-connected", and therefore eligible for compensation. As shown in Figure 1, there are three main rule conditions that a veteran must prove (connected to the ultimate conclusion at the top by the logical connective "AND"), and within each branch there are specific conditions if the claim is for PTSD (connected to the branch by "OR", indicating that alternative disabilities may have their own particular rules). In a BVA decision on such a disability claim, therefore, we expect the fact-finding reasoning to be organized around arguments and reasoning on these three PTSD rule conditions. Therefore, the rule tree governing a legal adjudication (such as a BVA case) provides the issues to be proved, and an organization structure for classifying arguments or reasoning based on the evidence. The critical connectors between the rule conditions of the rule tree and the evidence in a specific case are the Finding Sentences.

3.2 The Qualitative Study Datasets

From the common dataset of 50 BVA decisions we randomly drew a set of 5 decisions to function as the **qualitative-study observation sample** ("**QS-OS**"). The QS-OS is the sample of labeled sentences that we studied qualitatively to hypothesize classification features for rhetorical roles. The QS-OS dataset contains 530 sentences, with the following frequencies for particular sentence roles: Finding Sentences = 58, Evidence Sentences = 201, Reasoning Sentences = 40, Legal-Rule Sentences = 81, Citation Sentences = 103, and other Sentences = 47.

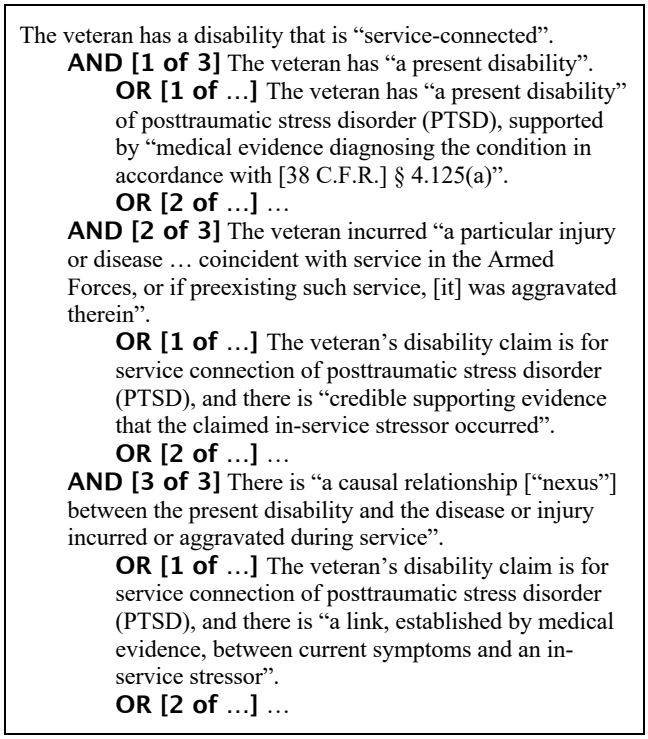

Figure 1. High-Level Rule Tree for Proving a Service-Connected Disability, and Specifically PTSD.

Theorists on sample size for qualitative studies have determined that the appropriate size depends upon many factors [24]. They recommend that researchers can stop adding to the observation sample once that sample has reached reasonable "saturation," such that it is sufficiently information-rich and adding more members would be redundant. [24, 12] In the present study, rather than devising a metric for saturation, we decided to test our main hypothesis by randomly drawing a very small sample of 5 decisions, analyzing the 58 sentences labeled as Finding Sentences in those decisions, forming hypotheses about predictive classification features, and testing the predictive power of those features.

The **qualitative-study test sample** ("**QS-TS**") consists of the remaining 45 BVA decisions from the PTSD dataset, excluding the 5 decisions we used to create the QS-OS dataset. As we formulated hypotheses about the classifying power of linguistic features based on the QS-OS, we tested those features quantitatively against the QS-TS. Within these 45 decisions, we isolated only the evidence assessment portions of the decisions, the extended section under the heading "Reasons and Bases" for the findings. We call this set of labeled sentences the "**QS-TS-R&B**". This dataset contains 5,422 sentences, with the following frequencies for particular sentence roles: Finding Sentences = 358, Evidence Sentences = 2,218, Reasoning Sentences = 669, Legal-Rule Sentences = 857, Citation Sentences = 1,015, and other Sentences = 305. We used QS-TS-R&B to test our observation-based hypotheses about predictive linguistic features.

3.3 The Machine Learning Dataset

For our ML experiments, we started with the entire PTSD-Sent dataset and performed certain preprocessing. We removed sentences that are merely headings, as well as numeric strings in the data. All words that remained were stemmed using NLTKs Snowball stemmer. Since punctuation symbols such as hyphens appear to interfere with the stemmer, we filtered out all non-alphabetic characters prior to the stemming step. If the filtering and stemming processes reduced a sentence to only blank characters, the entire sentence was dropped. Importantly, English stop words were not eliminated. Considering that each instance is a relatively short English sentence, eliminating any words might increase the classification error rate.

This preprocessing stage reduced the total data set to 5,797 usable labeled sentences. The frequencies of sentence types after preprocessing were: Finding Sentences = 490, Evidence Sentences = 2,419, Reasoning Sentences = 710, Legal-Rule Sentences = 938, Citation Sentences = 899, and other Sentences = 341.

The features chosen for the machine learning algorithm were the individual tokens in all the sentences (3,476), and the bigrams (30,959) and trigrams (59,373) that appear in them. These features also form the vocabulary for the vectorizer. We used the CountVectorizer class of the Scikit-learn Machine Learning library [25] as the feature extractor. The size of the vector was equal to the vocabulary size (93,808). On average, each sentence had about 60 true entries.

4 Results of the Qualitative Study

This Section describes the experiments we conducted in the qualitative study, as well as the results of those experiments. As we discussed in Section 3.2, the qualitative study was designed to test our main hypothesis that we can use a very small observational sample (only 5 decisions, containing 530 labeled sentences) to develop classifying scripts that perform reasonably well against the remainder of the PTSD dataset (a test dataset of 45 decisions, containing 5,422 labeled sentences), at least for purposes of some use cases. We also use the qualitative study to test our second hypothesis that attribution theory provides a general and transferable method for creating semantic types and linguistic features.

4.1 The Qualitative Study Methodology

In order to develop a systematic methodology for discovering linguistic features that might classify Finding Sentences, we used attribution theory. An example of a sentence explicitly stating an attribution relation is: *The Board finds that the veteran currently has PTSD*. In interpreting the meaning of this sentence, we attribute to "the Board" the conclusion that "the veteran currently has PTSD". As illustrated in this example, attribution relations have at least three elements or predicate arguments [22, 41]: (A) the **attribution cue** that signals an attribution, and which provides the lexical grounds for making the attribution (in the example, *finds that*); (B) the **attribution subject**, or the actor to which we attribute the propositional content of the sentence (in the example, the Board); and (C) the **attribution object**, or the propositional content that we attribute to the attribution subject, expressed in normal form by an embedded clause (in the example, *the veteran currently has PTSD*). We distinguish the attribution cues and attribution subjects, on the one hand, from the proposition being attributed. We call the former "finding-attribution cues" because a lawyer uses them to determine whether a sentence states a finding of fact or not, regardless of which legal-rule condition might be at issue. The proposition being attributed, on the other hand, is the content of the finding. In the example above, the finding-attribution cues are "*The Board finds that*", while the attribution object is the proposition "the veteran currently has PTSD." An important reason for separating these two categories and testing their performance independently is that finding-attribution cues are more likely to be transferable to disabilities other than PTSD, and they are more likely to have counterparts even in other areas of law.

4.2 Experiments with Finding-Attribution Cues

We conducted a qualitative study of the finding-attribution cues that occur within QS-OS, and ran various experiments to determine how scripts built on those cues would perform against QS-TS-R&B. This section reports the results of several of those experiments, with the results tabulated in Table 2.

4.2.1 Experiments E1 and E1N

It appeared from the QS-OS that a highly-predictive single word might be "finds". Although in this experiment we did not perform part-of-speech tagging, the word "finds" is generally used as a main verb (present tense, singular) when the Board states a finding. This is contrasted with Evidence Sentences, in which the verb is generally in the past tense (e.g., "found"), and the sentence attributes a proposition to a witness or document in the evidentiary record. We also observed occurrences of "concludes" and "grants" used in the same way as "finds". We ran these three alternatives as a single experiment, using the regular expression (finds | concludes | grants), with the results shown as **E1** in Table 2.

As shown in Table 2, a common mis-classification in E1 was with Legal-Rule Sentences. In Section 4.3 below, we discuss why precision is important for our use cases. By examining the Legal-Rule Sentences in QS-OS, we noted that, consistent with our main hypothesis, certain types of words and phrases occur in those sentences that we use to attribute them to legal authorities as sources of general legal rules. Such words and phrases include indefinite noun phrases (such as "a veteran," as contrasted with "the Veteran"), conditional terms (such as "if" and "when"), and words typically used as cues for attributing propositions to higher courts (such as "held that" or "ruled that"). We tested scripts that used such words or phrases to exclude Legal-Rule Sentences from the results of E1, with the results shown in Table 2 for **E1N**.

4.2.2 Experiments E2 and E2N

A primary strength of a qualitative study is being able to identify a phrase that might be highly predictive of Finding Sentences due to the legal meaning of the phrase. One such phrase is "preponderance of the evidence", which is used to formulate the legal standard for

finding a proposition to be a fact. An alternative phrase that is often used when assessing what the total evidence proves is "weight of the evidence". We ran scripts using these two alternatives against QS-TS-R&B, with the results shown in Table 2 as **E2**.

	E1	E1N	E2	E2N	E1+2	E1N+2N
Finding	129	129	46	43	159	156
Evidence	3	3	0	0	3	3
Reasoning	67	66	2	2	69	68
Legal-Rule	14	10	18	2	30	12
Citation	0	0	0	0	0	0
Other	1	1	1	1	2	2
Recall	0.360	0.360	0.128	0.120	0.444	0.436
Precision	0.603	0.617	0.687	0.896	0.605	0.647
F1	0.450	0.455	0.216	0.212	0.512	0.521

Table 2. Qualitative Study Test Results (Frequencies) for Finding-Attribution Cues, by Sentence Rhetorical Role

As with experiment E1 above, the mis-classified Legal-Rule Sentences had the undesirable effect of lowering the precision of the script. By examining the Legal-Rule Sentences in QS-OS, we hypothesized that modal words or phrases, in addition to those indefinite, conditional and attributional words and phrases discussed in Section 4.2.1, could be used to exclude Legal-Rule Sentences. Examples of such modal phrases are "must determine" and "are not necessary." Scripts including these four types of words produced the results shown in Table 2 for **E2N**.

4.2.3 Experiments E1+2 and E1N+2N

In order to test a combination of scripts, we ran a script that classified a sentence as a Finding Sentence if either E1 so classified it or E2 did so. The results are shown as **E1+2** in Table 2. We also ran a combined experiment, including the Legal-Rule Sentence exclusion scripts from E1 (E1N) and from E2 (E2N), with the results shown as **E1N+2N** in Table 2.

4.3 Discussion of the Qualitative Study

We emphasize that we had a very limited objective in these experiments: to test, in a preliminary way, whether we could use attribution theory to develop hand-crafted, rule-based scripts that could perform adequately in a variety of important use cases. If we could observe useful linguistic patterns in only 5 decisions, we might be able to develop a general methodology that would be transferable to adjudicatory decisions in many areas of law.

We also stress that whether performance is adequate is a function of the end use case. For example, if the use case is to retrieve similar cases and to highlight sentences by rhetorical type for the purpose of suggesting how similar evidence has been argued in past cases, then the priority might be on precision over recall. This is because wasting the user's time with non-responsive returns might have a more serious cost than merely failing to retrieve all similar cases. For such a use case, even recall = 0.436 (for E1N+2N, Table 2) might be useful because nearly half of all Finding Sentences were correctly identified (true positives).

In addition, precision = 0.647 (for E1N+2N, Table 2) might be acceptable, because the false positives (sentences incorrectly classified as Finding Sentences) constituted only about 1/3 of the predicted sentences. Moreover, the largest number of mis-classified sentences occurred in Reasoning Sentences (68). This may be because a judge might use a main verb such as "finds" when reporting the Board's intermediate reasoning about the credibility or persuasiveness of individual items of evidence. Of the incorrectly classified sentences, about 80% were Reasoning Sentences, which are probably also instructive to a user who is looking for examples of arguments about evidence. For such a use case, a user might learn as much or more from reviewing a Reasoning Sentence as from reviewing a Finding Sentence, and confusion between these two rhetorical roles is less important. For these use cases (semantic search and semantic viewer), the performance of even these simple scripts could be very useful.

Contrast such use cases with a use case that calculates a probability of success for an argument pattern, based on historic results in decided cases. For such a use case, the validity of the probability would depend critically upon relative frequency in the database, and on high recall and precision of similar arguments from past cases. Retrieving every similar case would be a priority with a potentially significant cost of error – e.g., reliance on an erroneous probability in deciding whether to bring or settle a new legal case. Moreover, confusion between Finding Sentences (which record whether an argument was successful or not) and any other rhetorical type of sentence could have significant consequences.

Because we based the script development for these experiments on attribution theory, as well as on general concepts used to increase precision, we expect this methodology to be transferable to other legal areas besides veterans' disability claims.

5 Results of the Machine Learning Study

This Section describes the experiments we conducted in the ML study, as well as the results of those experiments. As described in Section 3.3, we filtered out certain sentences from the dataset, and stemmed the words, leaving us with a preprocessed dataset of 5,797 labeled sentences. Our goal was two-fold: to assess how well the chosen machine learning classifiers perform relative to each other and to the qualitative-study scripts; and to find out which features were determined by each classifier as being significant to the prediction of Finding Sentences. The algorithms we chose for this study are Naive Bayes (NB), Logistic Regression (LR), and support vector machines (SVM) with a linear kernel [26, 9, 8].

We ran each ML algorithm 10 times, each run using a randomly chosen training subset that contained 90% of the labeled sentences. The trained classifier was used to predict the labels for the remaining 10% of sentences. All results shown in this section are the averages over these 10 runs.

For each ML algorithm, we ran two sets of experiments. In the first set of experiments (the "multi-class" experiments) we retained the labels for all 5 sentence types in the PTSD-Sent dataset – i.e., each classifier was fit to a multi-class training set. We recorded the overall accuracy score (the fraction of correctly labeled test instances), the classification summary, and the confusion matrix for

each algorithm and each run. The classification summary records the precision, recall and F1-score for each label. A confusion matrix cell-value C[*i*][*j*] is the number of test sentences that are known to be in class *i* (row *i*) but are predicted by the classifier to be in class *j* (column *j*). All values shown are averages over the 10 runs.

In the second set of experiments (the "two-class" experiment), we labeled all sentences other than Finding Sentences as "Non-Finding" sentences, so the training and test datasets then contained only two classes. As before, we recorded the accuracy scores, the classification summaries, and the confusion matrices and averaged them over the runs of each algorithm. In addition, for the LR and SVM classifiers, we extracted the top 20 features as measured by their weights in the fitted classifier. Note that since Finding Sentences form only about 8.5% of the dataset, the default classifier that labels all test instances as "Non-Finding" would have an accuracy score of 91.5% (under reasonable assumptions about the distribution of sentences).

Table 4 summarizes the average accuracy for each classifier, for each of the two sets of experiments. We also computed the false positive rates from the confusion matrices. The remainder of this section of the paper reports details on each ML classifier.

Algorithm / Metrics	Multi-class Accuracy	Multi-class False-Pos	Two-class Accuracy	Two-class False-Pos
NB	81.7%	1.5%	93.4%	2.4%
LR	85.7%	1.6%	96.3%	1.2%
SVM	85.7%	1.6%	96.8%	1.2%

Table 4. Average Accuracy and False-Positive Rates, Three Classifiers, Two Sets of Experiments

5.1 Naive Bayes (NB)

The Scikit-learn Python module has implementations of multiple variants of the basic NB algorithm. We chose the GaussianNB implementation with default parameters to present results (implementation of ComplementNB yielded similar results). Results for the two-class experiment are shown in Table 5.

	Precision	Recall	F-1
Finding	0.64	0.48	0.54
Non-Finding	0.95	0.98	0.96

Table 5. Naive Bayes Classification Summary, Two-Class

Discussion: The results show that NB is not a preferable classifier for this problem. While the overall accuracy for both the multi-class and two-class experiments appear to be acceptable (Table 4), a closer look indicates these are substantial deficiencies in this classifier, especially for the important two-class case (Table 5). The two-class accuracy score of 93.4% (Table 4) is not a significant improvement over the default classifier (with an accuracy of 91.5%). The precision of 0.64 for Finding Sentences indicates that the classifier is likely to generate a number of false positives. The underlying issue is likely to be the strong assumption of conditional independence between the features. Finally, the inability of this model to indicate which features were most important in making the determination of Finding Sentences makes it an opaque classifier.

5.2 Logistic Regression (LR)

The LR algorithm produces a binary classifier, also known as a log-linear classifier. Since the LR algorithm produces only binary classifiers, for our multi-class experiments we used the one-versus-the-rest approach. Results are shown in Tables 6 – 8.

Discussion: The results show that LR is an acceptable classifier for this problem. The two-class accuracy score of 96.3% (Table 4) is better than that of the default classifier, although in this classifier as well most of the accuracy score appears to come from the correct predictions of the Non-Finding Sentences. The two-class precision of 0.84 for Finding Sentences (Table 8) indicates that false positives are still a concern, though substantially lower than those of the NB classifier. The confusion matrix did not indicate any dominant source of error. The words and phrases (stemmed) in the highest-ranked features were similar to those used in the hand-scripted classifier.

	Precision	Recall	F1-score
Citation	0.99	0.97	0.98
Evidence	0.87	0.94	0.91
Finding	0.81	0.78	0.79
Legal-Rule	0.88	0.91	0.89
Reasoning	0.66	0.52	0.58
Others	0.70	0.59	0.64

Table 6. Logistic Regression Summary, Multi-Class

	C	E	F	L	R	O
C	91.1	0.6	0.0	1.5	0.0	0.3
E	0.7	226.6	1.3	1.1	9.8	1.1
F	0.0	3.1	37.7	1.5	4.5	1.4
L	0.2	1.9	1.3	85.1	3.4	1.5
R	0.2	21.4	4.5	4.7	37.4	3.8
O	0.2	6.2	1.8	3.3	1.7	19.1

Table 7. Logistic Regression Confusion Matrix, Multi-Class

	Precision	Recall	F-1 Score
Finding	0.84	0.69	0.75
Non-Finding	0.97	0.99	0.98

Table 8. Logistic Regression Summary, Two-Class

5.3 Support Vector Machines (SVM)

An SVM is an ML algorithm for binary classification problems. It is based on finding a maximum margin hyperplane that divides the training set into the two classes. Based on the success of the LR classifier, we decided to use a linear kernel for the SVM. Since SVM classifiers are by default binary, for the multi-class experiment the implementation builds one-versus-one classifiers and a voting scheme is used to predict the label for a test instance. Some results are shown in Tables 9 and 10.

Discussion: The results show that performance of the SVM classifier with a linear kernel has similar performance to that of the LR classifier. This is true for both the multi-class and the two-class experiments. However, there is substantial divergence in the top features chosen by the two algorithms. The features in common are "board find", "thus" and "whether". One hypothesis is that most of the top features are used to decide the Non-Finding class labels, and the Finding class arises as a default class. Several of the highest-ranked features seemed to be specific for PTSD cases. Also, as with the LR classifier, the confusion matrix for the multi-class SVM did not indicate any dominant source of classification error.

	Precision	Recall	F1-score
Citation	0.98	0.96	0.98
Evidence	0.88	0.94	0.91
Finding	0.82	0.78	0.8
Legal-Rule	0.90	0.90	0.90
Reasoning	0.65	0.53	0.58
Sentence	0.63	0.63	0.62

Table 9. SVM Classification Summary, Multi-Class

	Precision	Recall	F-1 Score
Finding	0.85	0.74	0.79
Non-Finding	0.98	0.99	0.98

Table 10. SVM Classification Summary, Two-Class

6 General Discussion and Future Work

The main hypothesis for our work was that Finding Sentences in legal decisions contain such regular and distinctive phrasing that scripts written on a very small sample, as well as ML models trained on larger but still relatively small samples, could perform sufficiently well for many valuable use cases. The results of our preliminary experiments indicate that this hypothesis was correct, for the reasons we began to discuss in Section 4.3.

In the qualitative study, we used attribution theory to identify possible classification features from a very small set of 5 decisions, and we tested our hypotheses on a larger set of 45 decisions. Using attribution-finding cues and other general concepts, we developed scripts that performed reasonably well for such use cases as semantic search and semantic viewer, for the purpose of retrieving examples of reasoning in similar cases. Given the generic nature of the scripts and the small sample of labeled decisions, there is reason to think that this methodology is transferable to other areas of law. We plan to test this hypothesis in our future work.

For the ML experiments, for each of 10 runs we employed 90% of 5,797 labeled sentences for training, and the other 10% for testing. While this quantity of training/testing data was 10 times the quantity of data used to construct the hand-crafted scripts, it is still a smaller dataset than those on which ML models are typically based. The LR and SVM classifiers produced similar recall, precision and F1 scores for classifying Finding Sentences, in both the multi-class and two-class experiments. Either significantly outperformed the hand-crafted scripts in these metrics. However, we emphasize that we did not try to optimize the scripts that we tested. Our goal at this stage was to develop and test a methodology for writing such scripts, and to determine whether even basic scripts could yield promising results for some use cases. A next step is to improve the performance of our scripts in those use cases. One approach will be to employ part-of-speech tagging of at least subjects and verbs, which may improve the predictive power of script features by distinguishing between attribution subjects and cues, on the one hand, and attribution objects on the other.

A second approach will be to use our qualitative methodology to write and test scripts for the other rhetorical roles. Our results here suggest, for example, that there are promising scripts for excluding many Legal-Rule Sentences from consideration as Finding Sentences. We think that scripts can be written for positively classifying Legal-Rule Sentences. For example, in addition to any lexical features, a Legal-Rule Sentence is generally followed immediately by a Citation Sentence (or by intervening other Legal-Rule Sentences, and then a Citation Sentence). Moreover, Citation Sentences have very particular content and are highly distinguishable. Attribution theory will also guide script development for classifying Evidence Sentences. Thus, a larger qualitative study may lead to better-performing scripts.

We also intend to combine high-performing scripts into a pipeline that also includes ML or DL (deep-learning) classifiers. Scripts can add new and legally-significant labels to sentences, which can then provide input features for ML or DL classifiers. Training ML or DL classifiers on data partially annotated by scripts may improve their performance.

7 Conclusion

We used attribution theory to develop a qualitative methodology for analyzing a very small sample of labeled sentences to create rule-based scripts that can classify sentences that state findings of fact ("Finding Sentences"). We compared the results of those scripts with the performance of standard ML algorithms trained and tested on a larger dataset, but one that is still a relatively small dataset by ML standards. Both of these experiments suggest that some access-to-justice use cases can be adequately addressed with very small quantities of labeled data, and at much lower cost than previously believed.

ACKNOWLEDGMENTS

We thank the Maurice A. Deane School of Law for its support for the Research Laboratory for Law, Logic and Technology.

REFERENCES

[1] L. Al-Abdulkarim, K. Atkinson and T. Bench-Capon. 2016. Statement Types in Legal Argument. In Legal Knowledge and Information Systems (JURIX 2016), Bex, F., and Villata, S., eds. IOS Press, 3-12.

[2] M. P. Allen. 2007. Significant Developments in Veterans Law (2004-2006) and What They Reveal about the U.S. Court of Appeals for Veterans Claims and the U.S. Court of Appeals for the Federal Circuit. *University of Michigan Journal of Law Reform* 40, 483-568. University of Michigan.

[3] K. D. Ashley and V. R. Walker. 2013. Toward Constructing Evidence-Based Legal Arguments Using Legal Decision Documents and Machine Learning. In Proceedings of the 14th International Conference on Artificial Intelligence and Law (ICAIL 2013). ACM, New York, NY, 176-180.

[4] K. D. Ashley and V. R. Walker. 2013. From Information Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study. In Legal Knowledge and Information Systems, Ashley, K. D., Ed. IOS Press, 29-38.

[5] A. Bansal, Z. Bu, B. Mishra, S. Wang, K. Ashley, and M. Grabmai. 2016. Document Ranking with Citation Information and Oversampling Sentence Classification in the LUIMA Framework. In Legal Knowledge and Information Systems (JURIX 2016), Bex, F., and Villata, S., eds. IOS Press, 33-42.

[6] Board of Veterans' Appeals, U.S. Department of Veterans Affairs. 2018. Annual Report, Fiscal Year 2018.

[7] H. Bunt, R. Prasad and A. Joshi. 2012. First steps towards an ISO standard for annotating discourse relations. In *Proceedings of the Joint ISA-7, SRSL-3, and I2MRT LREC 2012 Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools* (Istanbul, Turkey, May 2012), 60-69.

[8] C. Cortes and V. Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 273-297. Kluwer.

[9] R-E. Fan, K-W. Chang, C-J. Hsieh, X-R. Wang and C-J. Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *J. Machine Learning Res.* 9, 1871-1874.

[10] M. Grabmair, K. D. Ashley, R. Chen, P. Sureshkumar, C. Wang, E. Nyberg and V. R. Walker. 2015. Introducing LUIMA: An Experiment in Legal Conceptual Retrieval of Vaccine Injury Decisions Using a UIMA Type System and Tools. In *Proceedings of the 15th International Conference on Artificial Intelligence & Law (ICAIL 2015)*, 69-78. ACM, New York.

[11] C. Grover, B. Hachey, I. Hughson and C. Korycinski. 2003. Automatic Summarization of Legal Documents. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL '03)*, 243-251. ACM, New York.

[12] T. C. Guetterman, T. Chang, M. DeJonckheere, T. Basu, E. Scruggs and V. G. Vinod Vydiswaran. 2018. Augmenting Qualitative Text Analysis with Natural Language Processing: Methodological Study. *J. Med. Internet Res*. 20(6), e231.

[13] B. Hachey and C. Grover. 2006. Extractive summarization of legal texts. *Artificial Intelligence and Law* 14, 305–345.

[14] R. Krestel, S. Bergler and R. Witte. 2008. Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the Sixth International Language Resources and Evaluation Conference* (LREC '08) (Marrakech, Morocco, May 28-30, 2008), 2823-2828.

[15] J. Lawrence and C. Reed. 2017. Mining Argumentative Structure from Natural Language Text Using Automatically Generated Premise-Conclusion Topic Models. In *Proceedings of the 4th Workshop on Argument Mining*, 39-48, Copenhagen, Denmark.

[16] E. de Maat, K. Krabben and R. Winkels. 2010. Machine Learning versus Knowledge Based Classification of Legal Texts. In *Proceedings of the 2010 Conference on Legal Knowledge and Information Systems* (JURIX 2010), 87-96.

[17] L. T. McCarty. 2007. Deep Semantic Interpretations of Legal Texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL '07)*, 217-224. ACM, New York.

[18] R. Mochales and M-F. Moens. 2011. Argumentation mining. *Artificial Intelligence and Law* 19, 1-22. Springer.

[19] M-F. Moens, E. Boiy, R. Mochales and C. Reed. 2007. Automatic Detection of Arguments in Legal Texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL '07)*, 225-230. ACM, New York.

[20] V. H. Moshiashwili. 2015. The Downfall of Auer Deference: Veterans Law at the Federal Circuit in 2014. *American University Law Review* 64, 1007-1087. American University.

[21] R. M. Palau and M-F Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, 98–107, Barcelona, Spain.

[22] S. Pareti. 2011. Annotating Attribution Relations and Their Features. In *Proceedings of the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR '11)* (Glasgow, Scotland, UK, October 28, 2011). ACM, New York.

[23] S. Pareti, T. O'Keefe, I. Konstas, J. R. Curran and I. Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, Washington, October 18-21, 2013), 989-999.

[24] M. Patton. 1990. *Qualitative Evaluation and Research Methods*. Beverly Hills, CA: Sage.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel and B. Thirion. 2011. Scikit-learn: Machine Learning in Python. *J. Machine Learning Res.*12, 2825-2830.

[26] S. E. Robertson and K. Spark Jones. 1976. Relevance Weighting of Search Terms. *J. American Society for Information Science* 27(3), 129-146.

[27] M. Saravanan and R. Ravindran. 2010. Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artificial Intelligence and Law*, 18, 45–76.

[28] J. Savelka, V. R. Walker, M. Grabmair and K. D. Ashley. 2017. Sentence Boundary Detection in Adjudicatory Decisions in the United States. *Revue TAL*, 58(2), 21-45.

[29] O. Shulayeva, A. Siddharthan and A. Wyner. 2017. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law* 25(1), 107-126.

[30] C. Stab and I. Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 46-56, Doha, Qatar.

[31] S. Teufel and M. Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4), 409-445.

[32] H. Wachsmuth, M. Potthast, K. Al-Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff and B. Stein. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining*, 49-59, Copenhagen, Denmark.

[33] V. R. Walker. 2014. Representing the use of rule-based presumptions in legal decision documents. *Law, Probability and Risk*, 13(3-4), 259-275. Oxford UP.

[34] V. R. Walker, P. Bagheri and A. J. Lauria. 2015. Argumentation Mining from Judicial Decisions: The Attribution Problem and the Need for Legal Discourse Models. Paper at the First Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts (ASAIL 2015), San Diego, California, USA. URL: https://people.hofstra.edu/vern_r_walker/WalkerEtAl-AttributionAndLegalDiscourseModels-ASAIL2015.pdf.

[35] V. R. Walker, N. Carie, C. C. DeWitt and E. Lesh. 2011. A framework for the extraction and modeling of fact-finding reasoning from legal decisions: lessons from the Vaccine/Injury Project Corpus. *Artificial Intelligence and Law* 19, 291-331.

[36] V. R. Walker, D. Foerster, J. M. Ponce and M. Rosen. 2018. Evidence Types, Credibility Factors, and Patterns or Soft Rules for Weighing Conflicting Evidence: Argument Mining in the Context of Legal Rules Governing Evidence Assessment. In *Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018)*, 68-78. ACL.

[37] V. R. Walker, J. H. Han, X. Ni and K. Yoseda. 2017. Semantic Types for Computational Legal Reasoning: Propositional Connectives and Sentence Roles in the Veterans' Claims Dataset. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL '17)*, 217-226. ACM, New York.

[38] V. R. Walker, A. Hemendinger, N. Okpara and T. Ahmed. 2017. Semantic Types for Decomposing Evidence Assessment in Decisions on Veterans' Disability Claims for PTSD. In *Proceedings of the Second Workshop on Automatic Semantic Analysis of Information in Legal Texts (ASAIL 2017)*, 10 pages, London, UK.

[39] B. Waltl, G. Bonczek, E. Scepankova and F. Matthes. 2019. Semantic types of legal norms in German laws: classification and analysis using local linear explanations. *Artificial Intelligence and Law* 27, 43-71. Springer.

[40] D. Walton. 2009. Argumentation theory: A very short introduction. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, 1-22. Springer, US.

[41] B. Webber and A. Joshi. 2012. Discourse Structure and Computation: Past, Present and Future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (Jeju, Republic of Korea, July 10, 2012), 42-54.