Evidence Types, Credibility Factors, and Patterns or Soft Rules for Weighing Conflicting Evidence: Argument Mining in the Context of Legal Rules Governing Evidence Assessment

Vern R. Walker, Dina Foerster, Julia Monica Ponce, Matthew Rosen
Research Laboratory for Law, Logic and Technology (LLT Lab)
Maurice A. Deane School of Law
Hofstra University; Hempstead, New York 11549, USA
Vern.R.Walker@Hofstra.edu

Abstract

This paper reports on the results of an empirical study of adjudicatory decisions about veterans' claims for disability benefits in the United States. It develops a typology of kinds of relevant evidence (argument premises) employed in cases, and it identifies factors that the tribunal considers when assessing the credibility or trustworthiness of individual items of evidence. It also reports on patterns or "soft rules" that the tribunal uses to comparatively weigh the probative value of conflicting evidence. These evidence types, credibility factors, and comparison patterns are developed to be inter-operable with legal rules governing the evidence assessment process in the U.S. This approach should be transferable to other legal and non-legal domains.

1 Introduction

Argument mining from the fact-finding portions of adjudicatory decisions in law presents several advantages. One advantage is analyzing careful reasoning by professional authors of unstructured natural language documents, which contain explanations of the reasoning of the decision maker from the evidence in the case to the factual conclusions. Another advantage is mining arguments and reasoning from documents that combine various types of evidence, such as lay testimony, expert opinions, medical records, and scientific publications. Yet another advantage is the societal importance of the subject matter, from disability claims to vaccine-injury compensation to medical malpractice.

If we can mine a large number of fact-specific cases for the arguments of the parties and the reasoning of the decision makers, then we could identify frequencies, trends, and success rates for different types of argument. We could also determine whether decision making among factually similar cases has been consistent. Native or web applications could use the data to recommend evidence and arguments to parties in new cases, and they could provide historically based support for legal decision makers.

But argument mining from adjudicatory decisions also faces significant challenges. Adjudicatory decisions occur within a complex legal process for resolving a dispute or deciding a case. In general, an adjudicatory process involves parties to the dispute, who raise issues to be decided, argue for or against specific outcomes on those issues, and often (especially in common law countries) produce the evidence on which the findings of fact are based. Another key participant in the adjudicatory process is the presiding official at the trial level (e.g., judge or administrative official), who presides over the creation of the official evidentiary record, decides which legal rules are applicable to the process, and decides how to enforce those legal rules. Another participant in the process is the fact-finder or trier of fact, who evaluates all the evidence produced, and officially declares the propositions that constitute the findings of fact for the proceeding. (Depending upon the tribunal and process, the same person may perform the roles of presiding official and of trier of fact.) In addition, there is almost always a reviewing authority (e.g., an appellate court), which oversees the decisions made at the trial level.

Such complexity of the adjudicatory process results in complex legal rules designed to govern the procedures. Substantive rules establish the issues to be decided, while process rules govern the procedures for deciding those issues (Walker, 2007). Some process rules govern various

Proceedings of the 5th Workshop on Argument Mining, pages 68–78

participants and actions related to assessing the probative value of the evidence. For example, such rules may govern the admissibility of evidence into the evidentiary record, dictate the relevance or irrelevance of certain types of evidence for particular conclusions, establish what some evidence presumptively proves, or determine when a set of evidence is minimally sufficient to warrant a reasonable inference. If argument mining is performed for the purpose of recommending possible arguments in new legal cases, then we must ensure that those arguments are constructed within the constraints imposed by such legal rules.

Such process complexity also results in decision documents from fact-finding tribunals in which sentences have a range of rhetorical roles (Walker et al., 2017a). Some sentences report the procedural history of the case, while others state the legal rules that are applicable to the case, or they provide citations to legal authorities. Other sentences may state the rulings of law on motions made by the parties, or they may explain the bases for those rulings. In mining the fact-finding reasoning from the decision, it is necessary to identify and exclude most of these types of sentences from those that contain the fact-finding reasoning. The tribunal's assessment of the evidence is expressed in sentences that summarize the evidence presented, that state the arguments of the parties based on that evidence, that state the findings of fact, and that explain the tribunal's inferences from that evidence to those findings.

This paper reports on research to empirically derive a typology for arguments that is flexible enough to type most evidence assessment actually found in adjudicatory decisions, and transferable to many substantive areas of law. The typology should have a reasonable likelihood of automatic and accurate classification, so that software can identify trends and success rates with acceptably low error rates, and software can make recommendations about arguments in new cases. The argument types should also be inter-operable with the complex legal rules in the U.S. that constrain the evidence assessment process.

This paper reports on a typology based on adjudicatory decisions about veterans' claims for disability benefits in the United States. Section 2 summarizes prior work relevant to our research. Section 3 describes the dataset and our methodology. Section 4 reports the types of evidence that the tribunal considers relevant to the major issues litigated in these cases. Section 5 reports, for two primary types of evidence, the factors that triers of fact consider in assessing the credibility or trustworthiness of individual items of that type of evidence. Section 6 reports on patterns or "soft rules" that the tribunal uses to comparatively weigh the probative value of conflicting evidence. Section 7 discusses the usefulness of this approach and future work.

2 Prior Work

Prior work in argument mining directly related to our project is work aimed at classifying arguments into types, and especially by means of classifying the premises of the arguments into types. Researchers generally identify a unit of argument as containing a conclusion or claim, together with a set of one or more premises. (E.g., Palau and Moens, 2009; Walton, 2009; Stab and Gurevych, 2014; Lawrence and Reed, 2017; Wachsmuth et al., 2017.) One approach to classifying arguments is using the argumentative relation between premises and conclusion (e.g., support, opposition; pro, con; stance) (e.g. Lawrence and Reed, 2017; Wachsmuth et al., 2017). Another approach is to classify arguments by types of premises. Our work uses both approaches, but this paper reports only on the latter approach.

As we explain in Section 3, we distinguish two kinds of premise (or correspondingly, two types of sentence or clause): propositions stating the relevant evidence and propositions stating the reasoning from that evidence to the conclusion (the finding of fact). Stab and Gurevych (2014) classified clauses as major claim, claim, premise or non-argumentative, with directed argumentative relations possibly running from a premise to a major claim, a claim, or another premise. Liebeck et al. (2016) adapted this approach to mining suggestions or claims on options for actions or decisions.

Boltužic and Šnajder (2016) developed a typology for premises organized along three dimensions: premise type (fact, value, or policy), complexity (atomic, implication, or complex), and acceptance (universal or claim-specific). Hidey et al. (2017) classified premises as logos, pathos, or ethos. They defined an "ethos" premise as one that "appeals to the credibility established by personal experience/expertise." As we explain in Section 4, it is common in legal evidence assessment to evaluate the credibility or trustworthiness of

sources of evidence. The adjudications in our dataset use a broad concept of credibility and a structured approach to determining credibility.

Some research directly related to our work is Rinott et al. (2015). They identified three different types of evidence: study results, expert testimony, and anecdotal descriptions. Also, Addawood and Bashir (2016) developed for Twitter data a typology of evidence having six types, including "expert opinion." We use somewhat related types, but we emphasize that our types are ones on which appellate courts have often predicated rules of law.

Finally, some research in argument mining has focused particularly on legal adjudicatory documents. Examples are U.S. court opinions (Jackson et al., 2003); judgments of the U.K. House of Lords (Hachey and Grover, 2006); U.S. trade secret cases (Ashley and Brüninghaus, 2009); extracts from legal texts of the European Court of Human Rights (Palau and Moens, 2009; Mochales and Moens, 2011); Indian court decisions (Saravanan and Ravindran, 2010); the process of argumentation from evidence to verdict in U.S. civil cases based on common law (Al-Abdulkarim et al., 2016); and judgment documents from Japanese civil trial courts (Yamada et al., 2017). However, to our knowledge, no research has developed a premise typology that is intended to be inter-operable with legal rules about evidence admissibility, irrelevance, minimal sufficiency, etc., particularly with such legal rules in the United States.

3 Dataset and Methodology

This section describes the sample of adjudicatory decisions we used in our study, as well as the methodology behind our results.

3.1 PTSD Dataset

We analyzed 30 fact-finding decisions issued by the U.S. Board of Veterans' Appeals ("BVA") from 2013 through 2016.[1] We arbitrarily selected those decisions from adjudicated disability claims by veterans for service-related post-traumatic stress disorder (PTSD). 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 (Allen, 2007; Moshiashwili, 2015). If the claimant is dissatisfied with the decision of the RO, she may file an appeal to the BVA. The BVA is an administrative appellate body that has the statutory authority to decide the facts of each case based on the evidence (Moshiashwili, 2015). 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).[2]

3.2 Methodology

For purposes of mining different types of fact-finding arguments or reasoning from these decisions, we first developed annotation protocols for identifying those sentences that generally do not contribute critical information. Examples of such sentences are those that primarily state the procedural facts of the case, the applicable legal rules, any rulings as a matter of law, or rationales for such rulings. We focus primarily on sentences that play one of three reasoning roles in evidence assessment: the **conclusion** (a **finding of fact**), which states whether a propositional condition of a legal rule is determined to be true, false or undecided; the **foundations** for the reasoning (the **evidence** in the legal record, such as the testimony of a lay witness, the opinion of an expert witness, or exhibits such as a medical record, a photo, or a published scientific study); and the **reasoning** from the evidence to the findings of fact. We call these, respectively, "**finding sentences**", "**evidence sentences**", and "**reasoning sentences**." We developed detailed annotation protocols for these three sentence types.

Table 1 reports the frequencies of occurrence for these sentence types in the PTSD dataset. Many decisions involve claims for multiple disabilities, of which PTSD is one. Table 1 provides the total number of sentences for entire decisions in the dataset. For the PTSD portions of those decisions,

[1] We cite decisions by their BVA citation number, e.g., "BVA 1400029." Decisions are available from the VA website: https://www.index.va.gov/search/va/bva.jsp.

[2] We follow the U.S. legal convention of citing to statutes, regulations, decisions, or other legal documents within the text.

it reports the frequency of evidence sentences, reasoning sentences, and finding sentences.

Semantic Type	Frequency
Sentence	8,149
Evidence Sentence	1,412
Reasoning Sentence	442
Finding Sentence	310

Table 1: Frequency of Sentences in Dataset, and of Sentence Types in PTSD Portions of Decisions

Next, we identified which PTSD-related finding sentences were relevant to determining the three major substantive issues of fact to be decided in these cases. To obtain compensation for a present disability, the veteran must prove that she has a disability that is "service-connected" (Walker et al., 2017b). This requires proving three major issues of fact: (1) the veteran has a present disability; (2) while in active service, the veteran incurred an injury or disease, or the veteran suffered an aggravation of a pre-existing injury or disease, or there occurred an "in-service stressor" that is capable of causing PTSD; and (3) there is a causal relationship (or "nexus") between the present disability and the in-service incurrence, aggravation, or stressor. We then identified the evidence sentences relevant to each issue, and we catalogued the types of evidence the BVA considered relevant to each major issue. We report some of our results in Section 4, with examples.

For each type of evidence that we found, we then searched our decisions for sentences that describe how the BVA evaluated the credibility or trustworthiness of a single item of such evidence. Such sentences are reasoning sentences. We found that normally these decisions refer to a number of factors that tend to increase or decrease the credibility or trustworthiness of the particular evidence. We report some of our results on these factors in Section 5, with examples.

We then searched for reasoning sentences that weighed the comparative probative value of conflicting evidence relevant to the same issue of fact. In Section 6, we report a few of the patterns that we are finding, with examples.

Throughout these searches (for evidence types, credibility factors, and patterns for comparing probative value), we noted legal rules on evidence assessment that govern the argumentation. Legal rules have the logical form of conditions – "if *p*, then *q*," where *p* states the condition of the rule and *q* its conclusion. We give examples of such rules in Sections 4, 5 and 6. We indicate in Section 7 how semantic typing of arguments should be inter-operable with governing legal rules.

4 Types of Evidence

We catalogued the types of evidence relied upon in the PTSD portions of the evidence assessment in the 30 BVA decisions. Table 2 lists the typology that we have developed. The Federal Rules of Evidence ("FRE") are typical of sets of rules adopted in U.S. jurisdictions to govern the admissibility of evidence in court. Broadly, evidence is admissible into the evidentiary record of a case, for consideration by the trier of fact, if it is "relevant," and not excluded by the U.S. Constitution, U.S. statutes, or rules of evidence. (Federal Rule of Evidence 402, 2017.) "Relevant evidence" is defined as evidence having "any tendency to make a fact [of consequence in determining the case] more or less probable than it would be without the evidence." (Federal Rule of Evidence 401, 2017.) In general, major categories of evidence are testimonial evidence (the testimony of a person), documentary evidence (evidence supplied by a writing or other document), and real evidence (physical evidence, such as clothing) (Black's Law Dictionary, 2014).

Evidence Type	Sub-Type
Lay Testimony	Veteran
	Veteran's spouse or partner
	Other veteran
	Other non-veteran
Medical Records	Pre-service
	In-service
	Post-service within the Veterans Administration
	Post-service not within Veterans Administration
Performance Evaluations	In-service
	Post-service
Other Service Records	
Other Expert Opinions	
Other Records	

Table 2: Types and Sub-Types of Evidence in the Sample of 30 BVA Decisions

Testimonial evidence in BVA cases includes lay testimony by the veteran filing the claim, a veteran's spouse or partner, another veteran, or other person. Such testimony often plays an important role in deciding the issues of fact. An example of an evidence sentence stating lay testimony is:

However, in written documents and in his Board hearing testimony, the Veteran contends that his acquired psychiatric disorder is the result of witnessing a shipmate fall overboard while stationed on the USS Constellation CVA 64 in approximately March 1975. [BVA 1554166]

Documentary evidence includes medical records, performance evaluations, and service or other records. Medical records, for example, can contain expert opinions, test results, or non-expert information. Examples of evidence sentences reporting the contents of medical records are:

With regard to positive evidence, in 2010 and 2011, a private psychologist, Dr. A.G., PhD., diagnosed the Veteran with PTSD due to his Vietnam experiences. [BVA 1400029]

The STRs [service treatment records] *showed no complaints, treatment, abnormalities or diagnosis for any psychiatric problems in service.* [BVA 1445540]

These examples of evidence sentences suggest the difficulty facing the BVA as trier of fact both in assessing the credibility of items of evidence taken individually (Section 5) and in resolving conflicts between items of evidence (Section 6).

5 Factors Affecting Credibility or Trustworthiness of Individual Items of Evidence

BVA decisions often take a factor-based approach to assessing the credibility of witness testimony or the trustworthiness of documentary evidence such as medical records. (For reasons we cannot discuss here, such factors are not identical to the issue-related "factors" common in the literature on case-based reasoning (e.g., Ashley and Brüninghaus, 2009; Al-Abdulkarim et al., 2016).) "Credibility" or "trustworthiness" denotes the quality of inspiring trust or belief. See, e.g., *Indiana Metal Products v. N.L.R.B.*, 442 F.2d 46, 51-52 (7th Cir. 1971). Appellate decisions often refer to "factors" as the characteristics of a witness that tend to make her believable, or the characteristics of testimony or a document that tend to make it worthy of belief. See, e.g., *Southall-Norman v. McDonald*, 28 Vet. App. 346, 355 (Ct. App. Vet. Cl. 2016).

5.1 Lay Testimony

Table 3 lists some of the factors that can affect the credibility of lay testimony. There are legal rules governing the assessment of evidence using such factors. Some legal rules determine the relevance or irrelevance of such factors to proving particular issues of fact. For example, an appellate court has held that in assessing the credibility of oral testimony, a hearing officer may properly consider "the demeanor of the witness, the facial plausibility of the testimony, and the consistency of the witness' testimony with other testimony and affidavits." *Caluza v. Brown*, 7 Vet. App. 498 (1995), *aff'd*, 78 F.3d 604 (Fed. Cir. 1996). Such rules, however, tend to identify some but not all relevant factors.

Aspects of Lay Testimony	Factors Affecting Credibility
Source of Testimony (Witness)	Demeanor of witness while testifying
	Character of witness
	Consistency of witness
	Bias, personal interest
Basis of Testimony	Degree of personal knowledge
	Awareness of other evidence
	Competence relative to content
Content of Testimony	Facial plausibility
	Consistency with other evidence
	Corroboration from other evidence

Table 3: Factors Affecting Credibility of Lay Testimony

The **first set of factors** that we found in our sample of BVA decisions includes characteristics related to **the source** of the testimony – the witness herself. Such characteristics include: the demeanor of the witness while testifying at a hearing; some aspect of the witness's character (e.g., being considered a "malingerer" while on active duty); the consistency of the witness's own statements

over time about the same events; or some bias of the witness, such as a personal interest in obtaining disability benefits at the time of the statement.

An example of a reasoning sentence that shows taking such factors into account is the following:

Additionally, the appellant has been extremely active in pursuing disability claims, including VA claims, EEOC claims, and workman's compensation claims. In the process of these claims, the appellant has provided testimony that is internally inconsistent and appears to frequently be shaped by the type of claim he is pursuing. [BVA 1413417]

The **second set of factors** concerns **the basis** for the testimony. These factors include: the degree of personal knowledge of the witness as the basis for the content of the testimony; the awareness by the witness of other evidence bearing on the testimony; and the competence of the witness to testify about the content. This last factor arises frequently in veterans' cases, as when the veteran seeks to testify about a psychological diagnosis but does not have the training to do so. An example is:

As a lay person, the Veteran is competent to report what comes to him through his senses, but he lacks the medical training and expertise to provide a complex medical opinion as to the etiology of an anxiety disorder. [BVA 1608262]

The **third set of factors** addresses **the content** of the testimony directly. Factors that affect credibility include: the plausibility (or implausibility) of the testimony on its face, such as internal inconsistencies; the extent of consistency or inconsistency with other evidence; and whether there is positive corroboration from other evidence. In addition to examples above, the following is an example of a reasoning sentence on corroboration:

Additionally, there are no medical records, police records, or changes in behavior that corroborate the Veteran's assertions. [BVA 1613894]

5.2 Medical Records

Table 4 lists some of the factors that affect the trustworthiness of information and expert opinions contained in medical records. There are legal rules governing evidence assessment, when it relies on such factors. For example, an appellate court has held that when assessing the probative value of documents for certain purposes, the BVA "may properly consider internal consistency, facial plausibility, and consistency with other evidence." *Caluza v. Brown*, 7 Vet. App. 498 (1995), *aff'd*, 78 F.3d 604 (Fed. Cir. 1996). However, such legal rules tend to state only some but not all relevant factors.

Aspects of Medical Records	Factors Affecting Trustworthiness
Source of Medical Record (author or source of content)	Relevant qualifications, expertise, etc.
	Bias, personal interest
Basis of Medical Record (e.g., physical examination, psychological evaluation)	Personal observation of patient
	Credibility or accuracy of other information relied upon
	Extent of patient's record taken into account
Content of Medical Record	Remarks that undermine conclusiveness
	Extent of detail
	Consistency with other evidence
	Corroboration from other evidence

Table 4: Factors Affecting Trustworthiness of Information within Medical Records

The **first set of factors** identified in our sample of BVA decisions includes characteristics related to **the author** of the medical record or **the source** of its content. Often, the author of the record is the same person who is asserting the proposition stated in the record. On occasion, however, the author states the assertion of some other person (e.g., of the veteran who recounts her medical history to a physician). Some factors are: the relevant qualifications, expertise, knowledge or skill of the medical or other expert to whom the content is attributed; or bias or personal interest of the expert. An example of a reasoning sentence is:

The Board finds that the clinician's opinion is competent because she is qualified through education, training, or experience to offer medical diagnoses, statements, or opinions. [BVA 1340434]

The **second set of factors** concerns **the basis** for the content of the medical record. Relevant factors for trustworthiness include: whether the content of

the medical record is based on observations made personally by the source; the credibility or accuracy of other information that the expert relied upon (e.g., reports of personal history by the veteran); and the extent of the patient's record that the expert took into account in creating the content of the medical record. Some example reasoning sentences are:

This examiner's opinion was rendered following review of the claims file and interview of the Veteran. No deficiency is found in either respect. The Board notes that the Veteran's interview took a "great deal of time." That is shown by the examiner's use of that phrase and the fact that the summarization of the interview is around three pages in length. There is no indication of reliance on an inaccurate factual premise in formulating the opinion. A rationale for it was offered by the examiner. It further is sufficiently thorough and detailed. [BVA 1303141]

The above examples also illustrate some factors from the **third set of factors**, which address **the content** of the medical record. These factors include: remarks in the content that might serve to undermine the conclusiveness of the content (e.g., "possibly"); the extent of detail provided in the content; the degree of consistency or inconsistency with other evidence; and whether there is positive corroboration from other evidence, especially other medical evidence. Another example is:

The relevant evidence on this question includes a February 2007 VA examination report where the examiner opined that it was at least as likely as not that the Veteran suffered PTSD in her youth and that this was "possibly aggravated" by service. [BVA 1343153]

6 Patterns or Soft Rules for Comparing Conflicting Items of Evidence

Our study disclosed that the BVA has also developed recurring patterns of reasoning when comparing the probative value of conflicting evidence on the same issue of fact. In this section we provide examples of such patterns.

6.1 Comparing Evidence of Same Type

The lists of factors presented in Section 5 also supply an analytic approach to comparing conflicting evidence of the same type (e.g., conflicting medical records). In BVA 1400029, for example, on the issue of diagnosis of present PTSD, there was a conflict between VA treatment records (declining to diagnose PTSD) and a medical record by a private psychologist (diagnosing PTSD). After evaluating each of the available medical records on relevant factors, the Board reasoned:

Overall, the November 2010 and September 2011 VA psychological examinations in particular were thorough, supported by an explanation, and considered the Veteran's history and relevant longitudinal complaints. The VA opinions and treatment records outweigh the private treatment of record on the issue of whether the Veteran has a PTSD diagnosis in accordance with DSM-IV.

Because the same factors apply to each item of the same type of evidence, those factors provide an analytic framework for comparing conflicting items.

6.2 Comparing Evidence of Different Types

Several patterns or soft rules have developed for comparing credible evidence of different types. An assessment pattern can function as a "soft rule" if a reviewing court has explicitly stated that it is reasonable for the BVA to follow such a pattern in its decisions. (See Walker, 2007.) A legal rule stating that a pattern of assessment is reasonable is a permissive rule, stating in effect that the BVA may follow such a pattern without great risk of reversal by a higher court. But such a rule does not mandate following that pattern – it merely officially recognizes the pattern. At some later time, an appellate court might make using such a pattern mandatory (a normal legal rule), especially after the pattern becomes generally used. We mention and illustrate several of those patterns or soft rules.

Incompetent lay testimony vs. competent medical opinion. An important credibility factor for lay testimony is competence relative to the content. While a veteran is competent to testify concerning her own symptoms, the Board is skeptical concerning the veteran's competence to testify concerning diagnosis or etiology. The appellate courts have held that lay persons (non-experts) may be competent to testify about medical diagnoses under some conditions, e.g., *Davidson v. Shinseki*, 581 F.3d 1313, 1316 (Fed. Cir. 2009). However, the Board can consider a difference in competence between a lay witness and medical

experts on an issue of diagnosis. In addition, the regulations on PTSD specifically require "medical evidence diagnosing the condition." Therefore, on the issue of diagnosis of present disability, we can find the Board deciding in accordance with the evidence in the medical records, despite testimony by the veteran to the contrary (e.g., BVA 1400029).

Lay testimony vs. contemporaneous documentary evidence. Given the possibility of bias or personal interest, together with the innate unreliability of human memory, the Board often favors contemporaneous documentary evidence over conflicting lay testimony made much later. While the appellate courts have held that the Board cannot automatically determine that competent lay testimony lacks credibility simply because it is uncorroborated by contemporaneous medical records, *Buchanan v. Nicholson*, 451 F.3d 1331, 1336 (Fed. Cir. 2006), the Board may weigh the absence of contemporaneous medical evidence against the lay testimony, *id.* at 1336-37. Thus, we find reasoning that states that contemporaneous documentary evidence outweighs later lay testimony, especially when the latter is discounted due to credibility factors. See, e.g., BVA 1340434.

The Benefit-of-the-Doubt Rule. Normally, the burden of proving the facts of a claim is on the party making the claim. When the probative value of supporting and opposing evidence is equally balanced, the party with the burden of proof must lose, as a matter of law. With veterans' claims, however, the statute places the burden of proof on the government, and it gives the benefit of the doubt to the veteran, 38 U.S.C.A. § 5107(b) (2018). Therefore, even when there is conflicting evidence of different types, if the Board considers the supporting and opposing evidence to be equally probative, then it must find the issue of fact for the veteran. E.g., BVA 1455333 (VA treatment records vs. VA examination record, on issue of present diagnosis of PTSD).

7 Discussion and Future Work

One important problem in argument mining is developing a useful typology for determining argument frequency, argument trends, or argument success rates. To predict argument outcomes in future cases, the argument typology should be based on the strength of the substantive evidence (the soundness of the argument), not merely on formal properties such as syllogistic form (the validity of the argument). In addition, a typology of arguments should be flexible enough to classify most evidence assessment actually found in adjudicatory decisions, and it should be transferable to many substantive areas of law. The typology should have a reasonable likelihood of automatic and accurate classification. Finally, for the use case of recommending arguments in actual cases, the arguments should comply with any applicable legal rules on evidence assessment.

Our future work includes developing such a typology for entire arguments, building upon the evidence types, credibility factors, and comparison patterns discussed in this paper. Evidence types provide a method not only for further classifying evidence sentences, but also for classifying arguments that rely in part on such evidence. Factors relevant to an evidence type provide independent variables for algorithms that predict the credibility or trustworthiness of a particular item of evidence. Patterns or soft rules for comparing different types of evidence provide methods for predicting the net probative value of a set of evidence that supports and opposes a conclusion on a particular substantive issue. Taken together, these layers of semantic classification provide a typology for arguments that identifies argument types in terms familiar to judges and lawyers, a methodology for predicting the strength of an argument, and a means of evaluating the status of an argument relative to any governing legal rules.

There are good reasons to think that our approach (evidence types, credibility factors, and comparison patterns) is sufficiently flexible, and transferable to areas of law outside veterans' claims. First, sets of rules such as the Federal Rules of Evidence govern (with some exceptions) many types of judicial cases, both criminal and civil (Federal Rule of Evidence 1101, 2017). Even where such rules of evidence admissibility do not govern (e.g., in many types of administrative adjudications), the basic conceptual categories from such rules are still employed. For example, distinguishing evidence into testimonial, documentary and real evidence is probably universal in the U.S., as well as classifying testimonial evidence into lay and expert. Moreover, even when the Federal Rules of Evidence do not apply, courts often consult them on questions that arise, such as what factors to consider in assessing the probative value of expert opinions (e.g., *Nieves-Rodriguez v. Peake*, 22 Vet.

App. 295 (2008)). Thus, our approach to classifying arguments by types of evidence should have broad applicability, even if a particular area of law tends to rely on different sub-types of evidence.

Second, appellate courts widely employ the concept of a "relevant factor" to establish rules governing the fact-finding process, in many different areas of law. (E.g., *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).) Moreover, the factors relevant to evaluating the credibility of a witness's testimony are general in nature (see the examples in Section 5), and we expect them to be applicable across most substantive areas within law.

Finally, we have already found some patterns of comparing evidence types in areas of law outside veterans' claims. For example, the pattern of favoring contemporaneous documentary evidence over conflicting, later lay testimony is a recognized pattern in decisions about compensation for injuries allegedly caused by vaccinations. (E.g., *Walton v. Secretary of the Department of Health and Human Services*, No. 04-503V, 2007 WL 1467307 (2007).) In sum, there are good reasons to conclude that our approach will be widely useful across many areas of law.

In order to generate a sufficient amount of valid semantic data using our typology, we also develop annotation protocols (classification criteria and methods) to give precise meaning to the semantic type, to train new annotators, and to review the accuracy of human annotations. In our protocols, attribution verbs are strong discourse indicators for evidence and finding sentences (see Stab and Gurevych, 2014), especially when combined with signature grammatical subjects, such as "the Veteran" or "the Board" (see Walker et al., 2015). We also use such protocols to guide the development of rule-based software and linguistic features for automatically classifying legal texts (e.g., Savelka et al., 2017). Stab and Gurevych (2014) have classified such features into 5 groups. 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 Lawrence and Reed (2017). The system architecture described by Rinott et al. (2015) for ranking candidates for context-dependent evidence might be promising in this regard. We will use the results of our present qualitative study as the framework for such future quantitative research.

The development of factors that increase or decrease either credibility or trustworthiness (Section 5) invites research into probabilistic models. Perhaps sufficient data could be obtained to develop models for predicting credibility and trustworthiness of evidence in new cases. If so, this approach would have applications beyond the legal domain.

Finally, we are developing techniques for achieving the inter-operability of the semantic types used for argument classification with the legal rules in the U.S. that constrain the evidence assessment process. We have provided examples of such legal rules throughout this paper, as well as elsewhere (Walker et al., 2017b). The semantic types used to classify arguments should include the concepts found in the conditions of such rules – e.g., rules governing "lay testimony," "expert opinion," or "medical record." Any analytic service that monitors or recommends arguments in legal cases would need to access and apply such legal rules.

The annotated dataset for this study will be publicly available on GitHub, at: https://github.com/LLTLab/VetClaims.

8 Conclusion

On the basis of an empirical study of veterans' disability claims in the United States, we are developing a typology for arguments. We think that our approach to evidence types, credibility factors, and patterns for comparing probative value will provide a useful typology for fact-finding arguments that is transferable to domains of law other than veterans' disability claims, and perhaps also to non-law domains.

Acknowledgements

We are grateful to the peer reviewers for this paper, whose comments led to significant improvements. This research was generously supported by the Maurice A. Deane School of Law at Hofstra University, New York, USA.

References

Aseel A. Addawood and Masooda N. Bashir. 2016. "What is Your Evidence?" A Study of Controversial Topics on Social Media. In

Proceedings of the 3rd Workshop on Argument Mining, pages 1-11. Berlin, Germany.

Latifa Al-Abdulkarim, Katie Atkinson, and Trevor Bench-Capon. 2016. Statement Types in Legal Argument. In Floris Bex and Serena Villata, editors, *Legal Knowledge and Information Systems (JURIX 2016)*, pages 3-12. IOS Press.

Michael 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.

Kevin D. Ashley and Stefanie Brüninghaus. 2009. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law* 17: 125-165.

Black's Law Dictionary. 2014. Bryan A. Garner, Editor in Chief. Thomson Reuters.

Filip Boltužic and Jan Šnajder. 2016. Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates. In *Proceedings of the 3rd Workshop on Argument Mining*, pages 124-133. Berlin, Germany.

Federal Rules of Evidence. 2017. 28 United States Code Annotated.

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14: 305-345.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan and Kathleen McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11-21, Copenhagen, Denmark.

Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell and Arun Vachher. 2003. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence* 150: 239–290.

John Lawrence and Chris 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*, pages 39-48, Copenhagen, Denmark.

Matthias Liebeck, Katharina Esau and Stefan Conrad. 2016. What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld. In *Proceedings of the 3rd Workshop on Argument Mining*, pages 144-153. Berlin, Germany.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22. Springer.

Victoria Hadfield Moshiashwili. 2015. The Downfall of Auer Deference: Veterans Law at the Federal Circuit in 2014. *American University Law Review* 64: 1007-1087. American University.

Raquel Mochales Palau and Marie-Francine 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)*, pages 98–107, Barcelona, Spain.

Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M. Khapra, Ehud Aharoni and Noam Slonim. 2015. Show Me Your Evidence – an Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440-450, Lisbon, Portugal.

M. Saravanan and B. Ravindran. 2010. Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment. *Artificial Intelligence and Law*, 18(1):45–76.

Jaromir Savelka, Vern R. Walker, Matthias Grabmair and Kevin D. Ashley. 2017. Sentence Boundary Detection in Adjudicatory Decisions in the United States. *Revue TAL*, 58(2): 21-45.

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

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff and Benno Stein. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49-59, Copenhagen, Denmark.

Vern R. Walker. 2007. A Default-Logic Paradigm for Legal Fact-Finding. *Jurimetrics* 47: 193-243.

Vern R. Walker, Parisa Bagheri and Andrew 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.

Vern R. Walker, Ji Hae Han, Xiang Ni and Kaneyasu Yoseda. 2017a. 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 2017)*, pages 217-26, London, UK.

Vern R. Walker, Ashtyn Hemendinger, Nneka Okpara and Tauseef Ahmed. 2017b. 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 Semantic Information in Legal Text (ASAIL 2017)*, London, UK.

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

Hiroaki Yamada, Simone Teufel and Takenobu Tokunaga. 2017. Annotation of argument structure in Japanese legal documents. In *Proceedings of the 4th Workshop on Argument Mining*, pages 22-31, Copenhagen, Denmark.