The Need for Annotated Corpora from Legal Documents, and for (Human) Protocols for Creating Them: The Attribution Problem

Vern R. Walker
Director, Research Laboratory for Law, Logic and Technology
Maurice A. Deane School of Law
Hofstra University, New York, USA
Vern.R.Walker@Hofstra.edu

Abstract

This paper argues that in order to make progress today in automating argumentation mining from legal documents, we have a critical need for two things. First, we need a sufficient supply of manually annotated corpora, as well as theoretical and experimental evidence that those annotated data are accurate. Second, we need protocols for effectively training people to perform the tasks and sub-tasks required to create those annotations. Such protocols are necessary not only for a team approach to annotation and for quality assurance of the finished annotations, but also for developing and testing software to assist humans in the process of annotation. Drawing upon the latest work at Hofstra University's Law, Logic and Technology Research Laboratory in New York, the paper offers an extended example from the problem of annotating attribution relations, as an illustration of why obtaining consistent and accurate annotations in law is extremely difficult, and of why protocols are necessary. Attribution is the problem of determining which actor believes, asserts, or relies upon the truth of a proposition as a premise or a conclusion of an argument. The paper illustrates that in applying argumentation mining to legal documents, annotating attribution relations correctly is a critical task.

Relevant Seminar Topics

- Automated identification of arguments
- Automated identification of relationships between arguments
- Argument annotation methods and tools
- Annotation of argumentation corpora
- Applications of argument mining to legal documents

1 Introduction: Some Peculiar Features of Artificial Intelligence and Law

This paper addresses the goal of developing software to detect and extract patterns of argument and reasoning (argumentation mining) from legal texts, for the purpose of assisting lawyers and litigants in constructing potential arguments for new situations or cases.

There is available today an exploding array of available analytics and other software tools, such as IBM's Bluemix and Google's open-source TensorFlow. There is also widespread and growing expertise available in natural language processing and language technologies. But application of these tools and this expertise to legal documents has fallen far, far behind similar applications in health care and bioinformatics, for example. What is the explanation for this phenomenon? After all, electronic databases of legal documents date back to the early 1980s (e.g., commercial databases such as Westlaw and LexisNexis), and legal documents are becoming freely accessible through the websites of governmental institutions.

Part of the explanation is that researchers in the legal academy have generally ignored the development of practical, operational theories of legal reasoning, which could be used to inform software developers about what they should be looking for in legal documents. Without knowing in sufficient, concrete detail what lawyers and judges "see" when they read and understand legal texts, there is no sound theory for what developers of semantic analytics should be searching for in texts, and no sound metrics for how close they come to finding what they should be finding. Without an adequate, operational theory of how lawyers understand the meaning of a le-

gal text, software developers have no sound model to simulate or emulate.

A second part of the explanation for the lack of application to legal texts is that the task appears particularly difficult and daunting using today's software tools. Law faculties educate law students in how to "reason like a lawyer," but it takes many years of study to develop even a basic proficiency in the needed skills. Moreover, even experienced lawyers often differ on how to analyze important legal texts – indeed, analyzing legal texts insightfully and creating convincing legal arguments about their interpretation is the core task of the legal profession. When one mentions research into "artificial intelligence and law" to a practicing lawyer, the immediate response is nearly always some form of: "You are trying to develop machines that can replace lawyers?" While there are many good responses to that voiced worry, the correct underlying instinct of most lawyers is that "legal reasoning" (whatever that is) is at the heart of what lawyers do, and there is substantial doubt about whether any software today can even begin to perform that task. Understanding the possible alternative meanings of the text of a legal document, appreciating their legal significance, and using that information to craft potential arguments for new cases, seem like particularly difficult tasks for artificial intelligence as we know it today.

Even with the advanced machine learning approaches that we have today, this core task of argumentation mining seems well beyond our capabilities. It is one thing (although certainly a significant thing!) to mine Wikipedia pages for text passages and information needed for performing question-answer tasks. And it is one thing (and an impressive thing!) for technology-assisted review (TAR) systems to search "big data" sets of millions of emails for documents of relevance to a query, and to identify for redaction those portions of text in relevant documents that might enjoy a legal privilege, in order to make lawyers marginally more efficient in document review. But it is quite another thing to mine judicial decisions for threads of argument and reasoning, and to extract information to answer queries about legal rules or legal policies or findings of fact, in order to formulate possible arguments for future cases. While some tasks that software is able to do at some level of proficiency can be useful in legal practice (for example, question-answer, e-discovery), the core of what a lawyer does (argument mining from legal documents) remains a goal over the research horizon.

A third part of the explanation is that reasoning and inference in law does indeed have some peculiar features. Legal reasoning is pragmatic in the sense of being oriented toward regulating action, of occurring in real time as a result of limited resources and based on incomplete information, and of balancing accuracy against competing non-epistemic objectives (Walker et al. 2015b). And especially in judicial proceedings, inferences are often made through a structured process in which multiple actors play various roles, and particular actors make decisions concerning various logical components of the overall inference (Walker et al. 2013). The software that will prove adequate to assist in legal argumentation and reasoning will have to respect the pragmatic aspects of such participatory processes. And it will have to be adept at extracting information about such argumentation processes from legal documents such as judicial decisions.

This paper does not discuss software solutions to these challenges. Instead, it discusses two major barriers today that are critically impeding the successful application of argumentation mining to legal documents. The first barrier is the lack of annotated corpora that can be used for software development and testing. The second barrier is the lack of protocols for training human annotators and for achieving quality assurance over their annotations. After discussing these two barriers in general, the paper illustrates the nature of these problems by discussing the annotation of attribution relations in judicial decisions as an extended example.

2 The Problem: The Need for Annotated Legal Corpora and (Human) Protocols

This section of the paper outlines in a qualitative way two major barriers to developing automated argumentation mining for legal documents.

2.1 The Need for Annotated Corpora from Legal Documents

When a lawyer reads a judicial decision, hunting for the elements of a potential legal argument, what is it that she "looks for" and "sees"? A practicing attorney is likely to respond that she looks for and sees such things as legal rules, presumptions of law, arguments based on legal principles or legal policies, information about the legal procedures involved in the case, allegations of the parties, testimony of the witnesses, opinions of the experts, reasoning of the experts, ba-

ses for the expert opinions, critiques of the credibility of witnesses, findings of fact, rulings on the law, and so forth. Such a list forms the "stuff" of legal education, and discourse between lawyers occurs on this conceptual level. As with any native or trained reader for content in a specialized field, a lawyer comprehends the meaning of a legal document largely without attending to the linguistic cues that are the means of identifying that meaning.

But software engineers who hope to design artificial systems that can assist with, participate in, simulate, or act in reliance on such a process of text comprehension need lawyers not only to identify what the lawyer is looking for in an abstract sense (e.g., a "legal rule"), but also to provide a sufficient number of annotated examples that can be used for training and testing.

Moreover, the annotated data must be reasonably valid or accurate, with respect to adequately capturing the true or intended meaning of the text. And if multiple annotators share the annotation task, then there must be reasonable inter-annotator agreement on how to annotate the text, before there is any hope for achieving a consistent and accurate set of annotated corpora.

Finally, the types of semantic information that a lawyer may find useful in argumentation mining might differ from those that are useful in filing a tax form, or in monitoring for regulatory compliance. What types are useful depends in part upon the ultimate objective, and success in mining information is necessarily a function of that objective. And one set of semantic types might not be suitable, or at least sufficient, for all purposes.

2.2 The Need for (Human) Protocols

I am using the phrase "human protocol" as shorthand for a method or process by which a human lawyer can perform a sub-task that is part of the activity of reading a legal document with comprehension. If the reading is done for the purpose of argumentation mining, for example, this means developing protocols by which lawyers could perform the sub-tasks that together constitute understanding the argument-relevant information in the text. We must fill the gap described in Section 2.1 – the gap between lawyers' comprehending the meaning of a legal text and their having an operational method for arriving at that comprehension.

First, developing such protocols requires decomposing the task of "intelligent reading" into sub-tasks. Until we break "reading intelligently" into many sub-parts, we cannot hope to develop software that will emulate or even assist human performance.

Second, for each sub-task, this approach requires developing protocols whose input consists of: (A) observable information from the document, usually linguistic in nature, and (B) background knowledge that the reader brings to the comprehension of the document. The output of executing the protocol would be text annotated for argument-related elements. Those annotations would need to include the semantic and pragmatic typing of various important elements.

Third, there should be methods for integrating the output from the various sub-tasks into the desired overall outcome. This is where the purpose behind the reading becomes particularly important. If the purpose is to extract information from statutes, regulations, and past judicial decisions and use that information to help formulate effective arguments in new legal cases, then such new arguments become the ultimate output of the human annotation.

Without protocols for how humans can perform the sub-tasks and integration with reasonable proficiency and accuracy, there is very little hope of developing the valid (or even reliable) annotated corpora discussed in Section 2.1. And without such protocols for humans, there is little hope of developing a sufficient supply of annotated corpora for machine learning.

Moreover, without such protocols for how humans correctly perform the sub-tasks and integration, software engineers will have very few intuitions about how to design software to accomplish a similar task on smaller datasets. This is not to suggest that software methods must track or simulate human methods. But human methods often lead to insights about how to structure the software code. Indeed, if, as I suspect, argumentation mining in many areas of law turn out to be "little data" problems instead of "big data" problems, then insights into methods of problem solving will be even more valuable.

3 An Extended Example: Annotating Attribution Relations

As an illustration of the two challenges or barriers discussed in Section 2 above (creating annotated legal corpora and human protocols), this section discusses a particularly difficult sub-task for human or machine reading of a judicial decision for purposes of argumentation mining. This sub-task is the identification and semantic anno-

tation of "attribution relations" within the text. The problem of attribution has a role in argumentation mining in many contexts other than law. But solving the attribution problem in a legal context poses perhaps the most significant challenges when it comes to producing annotated corpora and human protocols.

This discussion is adapted from (Walker et al. 2015a). It is based on our work at the Hofstra University Research Laboratory for Law, Logic and Technology (LLT Lab), which I direct. We currently have projects for annotating argumentation in three types of judicial decisions in the United States: vaccine-injury compensation decisions, medical malpractice decisions, and decisions about veterans claims.

3.1 The Attribution Problem

In the context of argumentation mining, the attribution problem is the descriptive sub-task of determining "who believes what" – that is, determining which actor or participant subscribes to the truth of, concurs with, or relies upon a stated proposition. When I refer to "belief" in this context, I do not mean simply the mental states of human actors, but also a more general relation within argumentation. The proponent of an argument believes, asserts, or relies upon the truth of a proposition as a premise or a conclusion of that argument. Attribution helps to answer the question: "Whose argument is it?"

Accurate attribution of propositions is often a critical task for argumentation mining. Unless a document relates only propositions that are believed, asserted or relied upon by a single actor and no one else, it becomes critical to determine who is asserting what. For example, a judge might write in her decision the sentence *the varicella vaccine can cause neuropathy in humans*, but writing this sentence does not always indicate that the judge herself believes the stated proposition to be true. The sentence might report the unproved allegation of a party in a legal pleading, or the testimony of an expert witness, or the text of a document exhibit, or a conclusion or finding of fact by the judge herself. Argumentation mining from judicial decisions requires accurate attribution of stated propositions to the parties, witnesses, and documents placed in evidence, as well as to the judge or factfinder.

Attribution within the context of argumentation mining from legal documents is merely a special case of a broader problem of mining attribution relations. Attribution, as a general type of relation, is a classic problem area in natural language processing (Bunt et al., 2012; Krestel et al., 2008; Pareti 2011; Pareti et al., 2013). As a clear example of the problem, if a sentence explicitly attributes a proposition to some source by using a direct quotation, then this is some evidence for attributing the content of the quotation to that source (Krestel et al., 2008). However, quotation generally does not imply that the author of the sentence attributing the quotation believes the content of the quotation (the proposition being attributed) to be true. Support (if any) for attribution of the quoted proposition to the sentence author often must derive from the context in which the sentence was written, rather than from the semantics of the sentence itself. If this is true for direct quotations, then attribution in non-quotation situations is generally even more complicated (Pareti et al., 2013).

When it comes to the attribution problem in respect to argumentation mining from legal documents, there has been very limited work. Grover et al. (2003) reported on a project to annotate sentences in House of Lords judgments for their argumentative roles. Two tasks were: (i) attributing statements to the Law Lord speaking about the case or to someone else (attribution); and (ii) classifying 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 Teufel and Moens (2002) on attribution in scientific articles. The House of Lords judgments studied by Grover et al. (2003) treated facts as already settled in the lower courts, and engaged in policy-based reasoning about issues of law. I have found no empirical work that focuses on the problem of attribution in the factfinding portions of judicial decisions. However, such decisions might utilize scientific evidence, and therefore work on attribution in factfinding decisions complements the work of both Grover et al. (2003) and Teufel and Moens (2002).

3.2 The Elements of Attribution Relations

We can represent **attribution relations** using at least three main elements or predicate arguments, which is consistent with Pareti (2011):

(A) The **attribution object**: the proposition that we attribute to some actor, and which we infer the actor believes, asserts, or relies upon;

(B) The **attribution subject**: the actor to whom we attribute belief in the attribution object; and

(C) The **attribution cue**: the lexical anchor or cue that signals the attribution, and which provides the evidentiary basis for making the attribution.

In short, an attribution object is a proposition, an attribution subject is an actor, and an attribution cue is some word, phrase, or other linguistic cue. The attribution cue functions as the linguistic evidence supporting an attribution relation (Webber and Joshi, 2012).

3.3 Producing Accurate Manual Annotations

In reading a judicial decision to understand who is espousing what line of argumentation or reasoning, performing the sub-task of determining attribution relations is critical. The sub-task of accurately attributing propositions may seem relatively straightforward when a single sentence contains values for all three elements of the attribution relation, especially when the author of the sentence is a subject of the attribution. For example, we can find values for all three elements within sentence (1), taken as a sentence within a judicial decision written by a judge acting as the factfinder in the case:

(1) The court agrees with the testimony of the petitioner's expert witness that the varicella vaccine can cause neuropathy in humans.

In this sentence we find sound linguistic evidence for using the clausal complement embedded in the sentence (i.e., *the varicella vaccine can cause neuropathy in humans*) as the attribution object or attributed propositional content. The sentence also provides evidence of two subjects to which the content is attributed (i.e., *the petitioner's expert witness* and *the court*). Finally, we find sufficient linguistic cues for making this attribution (i.e., *the testimony of* for the petitioner's expert, and *agrees with* for the court).

However, it is very often the case that in judicial decisions we do not find plausible values for all three elements within a single sentence. Most often, we must resort to extra-sentence linguistic cues and presuppositional information to formulate and test hypotheses about attribution relations. Consider, for example, the following sentence (an embedded proposition within sentence (1)):

(2) The petitioner's expert witness testified that the varicella vaccine can cause neuropathy in humans.

The mere occurrence of such a sentence is generally not sufficient evidence that the judge herself believed that what the witness stated is true. The judge's sentence might simply be restating the witness's testimony. However, in a particular context, there might be warrant for attributing the proposition not only to the petitioner's expert witness, but also to the factfinder herself. For example, sentence (2) might occur in a paragraph where it is clear that the factfinder finds the petitioner's expert witness more credible than the opposing witness, the petitioner has no alternative witness or evidence on which the factfinder might rely, and the factfinder is in fact making an ultimate finding in favor of the petitioner. Given this context, we could reasonably infer that the judge as factfinder is also adopting or relying upon this testimony by the petitioner's expert.

Even the occurrence of the following declarative sentence (embedded in sentence (2)) might, in the right context, be sufficient evidence that the judge as author of the decision believes the stated proposition to be true:

(3) The varicella vaccine can cause neuropathy in humans.

For example, the occurrence of sentence (3) in a section of the decision entitled *Findings of the Court* might warrant attributing it to the judge as factfinder (but it does not warrant this inference in every case).

The previous paragraphs illustrate the subtleties of inference required for making accurate attributions within a judicial decision. But that discussion focused on three declarative sentences with straightforward syntactic structures. English sentences, however, can take diverse grammatical forms while still providing attribution information. Consider the following sentences drawn from actual judicial decisions, each of which provides attribution information:

(4) Her June 2001 symptoms of palpitations, rapid heart beat, shortness of breath, and diaphoresis were also consistent with myocarditis, in his opinion.

(5) The sort of post hoc ergo propter hoc reasoning offered by Dr. Corbier has been consistently rejected by the Court.

(6) Although Dr. Kane initially thought that Will's cerebellar ataxia was due to a viral infection, he changed his mind when he realized that the course of Will's ataxia

was lasting too long to be of the run of the mill type.

(7) The undersigned notes that the Federal Circuit in *Althen* affirmed the Honorable Susan G. Braden's decision that tetanus toxoid caused petitioner's optic neuritis.

Note just a few of the complexities for attribution on the basis of these sentences. Sentence (4) illustrates how solving the attribution problem is often dependent on solving the coreference problem (*his*). Sentence (5) illustrates the use of noun phrases to refer to propositional objects presumably expressed in some other sentence or sentences of the document, and illustrates other difficulties in identifying the precise propositional content to attribute to the Court. Sentence (6) illustrates the efficiency with which English can state multiple attributions in one sentence. Sentence (7) illustrates the nesting of attributions, but also raises such questions as whether the verb *affirmed* when predicated of an appellate court really means that we can attribute to the appellate court the finding of fact of the lower court.

With these examples, I have merely hinted at the extent of variability in legal texts when it comes to providing information about attribution. Express attribution can take a huge variety of syntactic forms, and implicit attribution is dominant in the style of English authors. This can lead to substantial inconsistency among human annotators. Law students, even after completing one or two years of legal education, produce results that are significantly unreliable (inconsistent) and inaccurate. Moreover, there is good reason to think that experienced attorneys would also display significant variability, since they have had no formal training in analyzing attribution relations.

The reality of low inter-annotator reliability and of inaccuracy for attribution relations introduces the next topic: the need to produce protocols for how humans could independently, consistently and accurately annotate attribution relations within legal documents.

3.4 Developing (Human) Protocols

The production of a sufficient quantity of accurately annotated corpora requires that multiple annotators be working on the project. In one sense, producing the protocols needed for each sub-task is a problem of effective education or pedagogy. However, it turns out to be a difficult problem to produce consistently accurate results. And perhaps surprisingly to someone not involved in legal education or who has not been through legal education, the legal academy has spent very little time trying to develop such pedagogical techniques, even though the skills involved seem central to legal reasoning.

Here I will give only one example of the problems encountered in developing protocols for identifying attribution relations. As noted in Section 3.3 above, identifying the appropriate attribution subjects (actors) and attribution objects (propositions) can pose significant problems, such as determining coreference and formulating precise propositions. But even if we adopt a strategy of setting aside for the moment the subtasks involved in solving these difficult reference problems, and focus attention on the sub-task of identifying and annotating attribution cues, we still encounter significant challenges for developing protocols. In general we can say that there are three types of problem. One type of problem is to determine which intra-sentence linguistic expressions can function as attribution cues. For some expressions the answer seems straightforward (e.g., *agrees with*), while for others the answer is more difficult (e.g., *affirms*). A second type of problem is to determine which extra-sentence but intra-document linguistic evidence can function as attribution cues (e.g., document segmentation). A third type of problem is identifying presuppositional or background information that is relevant in warranting attribution (e.g., the role of appellate courts with respect to trial courts).

These layers of problems suggest that decomposition into sub-tasks and developing protocols for specific sub-tasks is the only promising strategy. Until we have some adequate theory and consensus on exactly which types of semantic data we should be annotating, and until we can decompose the annotation process into sub-tasks and have developed and tested human protocols for performing and integrating those sub-tasks, we will have little hope for automating the annotation process. I also believe that progress down that road will be made by stages, using an approach parallel to that for TAR: as protocols for human annotators are being developed and tested for individual sub-tasks, we should be developing software that can assist (not replace) humans in performing that particular annotation sub-task. Gradually we can develop more powerful software analytics.

4 Conclusion

I have argued that in order to make advances in automating argumentation mining in law, we have a critical need for legal documents that have been manually annotated as corpora. Moreover, we must have theoretical and experimental reasons for regarding those annotated data as sufficiently accurate. In order to generate such corpora, we should decompose broader tasks into sub-tasks, and should develop protocols for effectively training people to perform those sub-tasks. Such protocols would be useful not only for performing quality assurance on the finished annotations, but also for developing and testing software to assist in the process of annotation. In such a way, we can hope to grow incrementally but steadily toward automated argumentation mining in law.

At the Hofstra University Research Laboratory for Law, Logic and Technology (LLT Lab), which I direct, we currently have projects for annotating argumentation in vaccine-injury compensation decisions, medical malpractice decisions, and decisions about veterans claims. Moreover, we are collaborating with Kevin Ashley and Jaromir Savelka at the University of Pittsburgh, and with Matthias Grabmair and Eric Nyberg at Carnegie Mellon University, in automating sub-tasks in text annotation. At the LLT Lab, we also develop, in addition to the semantic data as such, the protocols for guiding humans in the annotation of texts, and we work on writing software rules that can assist in following such protocols. It is painstaking work, but I believe it is necessary work in the right direction.

Reference

Bunt, H., Prasad, R., and Joshi, A. 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.

Grover, C., Hachey, B., Hughson, I., and Korycinski, C. 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.

Krestel, R., Bergler, S., and Witte, R. 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.

Pareti, S. 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.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. 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.

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

Walker, V.R., Park, C.H., Hwang, P.H., John, A., Krasnov, E.I., and Langlais, K. 2013. A process approach to inferences of causation: empirical research from vaccine cases in the USA. *Law, Probability and Risk*, 12: 189-205.

Walker, V.R., Bagheri, P., and Lauria, A.J. 2015a. Argumentation Mining from Judicial Decisions: The Attribution Problem and the Need for Legal Discourse Models. Presented at *ICAIL 2015 Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, June 12, 2015, San Diego, CA.

Walker, V.R., Lopez, B.C., Rutchik, M.T., and Agris, J.L. 2015b. Representing the Logic of Statutory Rules in the United States. Chapter in *Logic and Legislation* (Michał Araszkiewicz and Krzysztof Płeszka, eds.), in the Springer Series "Legisprudence Library" (Luc Wintgens and Daniel Olivier-Lalana, eds.).

Webber, B., and Joshi, A. 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.