Artif Intell Law (2017) 25:107–126
DOI 10.1007/s10506-017-9197-6

Recognizing cited facts and principles in legal judgements

Olga Shulayeva[1] · **Advaith Siddharthan**[1] · **Adam Wyner**[1]

Published online: 11 March 2017

Abstract In common law jurisdictions, legal professionals cite facts and legal principles from precedent cases to support their arguments before the court for their intended outcome in a current case. This practice stems from the doctrine of *stare decisis*, where cases that have similar facts should receive similar decisions with respect to the principles. It is essential for legal professionals to identify such facts and principles in precedent cases, though this is a highly time intensive task. In this paper, we present studies that demonstrate that human annotators can achieve reasonable agreement on which sentences in legal judgements contain cited facts and principles (respectively, $\kappa = 0.65$ and $\kappa = 0.95$ for inter- and intra-annotator agreement). We further demonstrate that it is feasible to automatically annotate sentences containing such legal facts and principles in a supervised machine learning framework based on linguistic features, reporting per category precision and recall figures of between 0.79 and 0.89 for classifying sentences in legal judgements as cited facts, principles or neither using a Bayesian classifier, with an overall κ of 0.72 with the human-annotated gold standard.

Keywords Legal judgements · Citations · Natural language processing

✉ Adam Wyner
azwyner@abdn.ac.uk

Olga Shulayeva
olga.shulayeva@abdn.ac.uk

Advaith Siddharthan
advaith@abdn.ac.uk

[1] Department of Computing Science, University of Aberdeen, Aberdeen, UK

1 Introduction

In common law jurisdictions, legal practitioners treat existing case decisions (precedents) as a source of law. Case citations, references to legal precedents, are an important argumentation tool, enabling lawyers to formulate and present their argument persuasively. This practice stems from the doctrine of *stare decisis*, which can be translated from Latin as to 'stand by the decided cases'[1], where a case under consideration that has facts similar enough to precedent cases should receive similar decisions as the precedents. A legal professional looks to establish the relevant law in the current case; to do so, she must consult precedent cases to establish what legal principles were applied to patterns of facts similar to the current case in order to decide the precedent. Citations from existing case law are used to illustrate legal principles and facts that define the conditions for application of legal principles in the current case.

Citation analysis can help legal practitioners to identify which principles have applied in a certain case and which facts have been selected as the 'material' facts of the case, i.e. the facts that influenced the decision and which are crucial in establishing the similarity between two cases. There is no defined guide on how to identify the law embedded within common law decisions, so legal professionals are expected to make themselves familiar with as many relevant decisions as possible in order to make informed predictions about the outcome of a current case. Decisions delivered by courts are binding and can therefore provide useful information for legal professionals. The information that is embedded within the cited cases includes the legal principles and facts that are used to reason to a decision. Optimally, a legal professional finds a cited case with the same facts and legal principles, and so can argue that the decision for the current case should be that of the precedent; similarly, the opposing party may identify precedents with opposing principles to argue the decision should be otherwise. More commonly, given that two cases are unlikely to be identical in every respect, legal professionals must consider a range of precedents, each of which highlight particular facts and legal principles that support their argument (or can be used to argue against the opposition). It is, then, essential that each side in the legal dispute identifies a relevant case base which supports the legal claims made during legal arguments. As the body of common law is continually growing, human citation analysis is complex as well as knowledge and time intensive.

To support citation analysis (discussed further in Sect. 3.1), existing electronic tools, such as electronic databases,[2] provide one word summaries for relationships between cases (e.g.'applied'). However, it is uncommon for them to extract information about the facts and the legal principles of the cited cases. This means that on many occasions readers are required to make themselves familiar with the full text of multiple law reports in order to identify the applicable law and the correct way to apply it. Thus, citation analysis tools save some labour by providing a preliminary filter on relevant cases, yet, identification of particular cases and the essential details require further manual effort.

[1] Source: http://thelawdictionary.org/.

[2] e.g. LexisNexis Shepard's Citations Service http://www.lexisnexis.com/en-us/products/shepards.page.

Before we describe our approach to citation analysis, certain key concepts of legal theory must be scoped, as this work is focused on the computational analysis of the language of the law rather than on legal theory. In particular, cases are considered to contain *ratio decidendi*, which can be translated as a *reason for a decision*, an important piece of reasoning that is incorporated into the argumentation structure of future decisions. A variety of approaches to defining *ratio decidendi* can be identified in legal theory. As defined by Raz (2002): "*ratio decidendi* can be identified as those statements of law which are based on the facts as found and upon which the decision is based". Greenawalt (2013) provides several explanations on what forms the binding part of a decision:

> (1) the rule(s) of law that the court explicitly states, or that can reasonably be inferred, that it regarded as necessary to (or important in) its resolution of the case [...], (2) facts the precedent court regarded as "material", i.e., crucial for the court's resolution, plus the result of the case; and (3) facts the court now constrained by the precedent regards as material in the earlier case plus its result.

The complexities stemming from the debates surrounding the definition of ratio are excluded from the scope of this paper, but we aim to annotate cited facts and principles, defined quite broadly to fit such definitions.

This paper makes a novel, preliminary contribution towards automated identification of legal principles and facts embedded within common law citations. A gold standard corpus is created, with sentences containing cited legal principles and facts manually annotated. A Bayesian Multinomial Classifier is then applied to the corpus using a set of linguistic features to automatically identify these sentences. The main results are a demonstration that (a) the human annotation task is feasible, i.e. human annotators can achieve reasonable agreement on which sentences in legal judgements contain cited facts and principles and (b) it is feasible to automatically annotate sentences containing such legal facts and principles to a high standard. The reported studies lay the basis for further applications, including creation of metadata for search and retrieval purposes, compilation of automated case treatment tables containing summaries about legal principles and material facts of cases, and automated analysis of reasoning patterns and consistency applied in legal argumentation.

We first present a motivation for our work in the context of legal research in Sect. 2. In Sect. 3, we then turn to related work. Then there are two studies, the first on manual annotation in Sect. 4 and the second on automated annotation in Sect. 5. The paper closes with some conclusions in Sect. 6.

2 Motivation in a legal context

It is expected that the automated analysis of citations can improve the compilation and use of *tertiary sources*, which are an essential tool for legal professionals. We develop this point in the context of legal research.

In order to build a successful argument lawyers need to identify applicable law which includes common law authorities. Legal principles that exist within the common law are formulated and encoded in existing decisions, so for lawyers it is often necessary to identify a set of factually similar cases which can be cited to support the argument. The research of legal authorities is often performed with the help of a number of tools, which according to Geist (2009) can be split into three categories—primary, secondary and tertiary:

- Primary sources are case law and legislation records that are created and distributed by a number of agents. An example of a primary source can be a common law report collated by a barrister.
- Secondary sources are materials containing commentary about case law and legislation (e.g. law journals).
- Tertiary sources are legal digests and citators that provide collections of systemised case law information.

Legal professionals usually rely on a combination of information from primary, secondary, and tertiary sources.

On many occasions the knowledge of legal discourse is important to under the legal principles that are encoded within primary sources.

Secondary sources, such as legal journals and textbooks, provide commentary collated by experienced lawyers and journalists, which familiarises the reader with the applicable law, correct ways of applying it, and debates surrounding it. Lawyers use law journals to simplify the search by employing the "expertise of jurists and journalists that follow legal matters closely" (Someren 2014).

Legal citators, which constitute tertiary sources, are systemised collections of law reports that facilitate a faster search of cases and provide brief case treatment summaries, indicating which cases were cited during the discussion. The first legal citators were printed on paper and their history date several centuries back from the modern era; see for example, overviews of the history of Anglo-American legal citation (Cooper 1982) and American legal citators (Ogden 1993a; Gerken 2016). Paper citators are still in circulation, but there are some disadvantages associated with their preparation and use. Firstly, compilation of case summaries requires humans to interpret the law, which can be time consuming, costly, and is often carried out companies providing citators as a service. The practice introduces and maintains a degree of opacity about the law as well as a dependency of legal practitioners on such citators. Secondly, paper citators are in constant need of updates since common law may be changed every time a new decision is made. Thirdly, manually searching through the ever growing database of the cases can be slow and not always efficient. Currently LexisNexis and WestLaw, large information service providers, offer electronic case citators. Such tools allow regular updates of the report databases and provide a Boolean toolkit, which enables a more efficient search. It is believed, however, that electronic citators still rely on humans to interpret the case law for case treatment summaries. It should also be noted that the results returned by the citators often require the reader to interpret and extract the applicable law that is encoded within the reports.

Though citators are an essential component in carrying out legal research, there are problems with current practice. It is expected that automated analysis of citations can allow making some improvements in compilation and use of tertiary sources.

Firstly, it should be noted that the improved means of information storage have lead to the information overload. The growing legal knowledge base makes it continually more difficult for the publishers to process and organise the information, as well as for the researchers to take the full benefit of the available data. For example, Mart (2013), who researched the efficiency of legal research performed with electronic citators, took a seed case *Regents v Bakke* and attempted to generate citation data for it. As reported by Mart, in Shepard's Citations, there were 6697 citing references and 1082 case citations for *Regents v Bakke*. In KeyCite (WestLaw), Bakke had 8882 citing references and 1031 case citations. Mart concluded that it would be hard for a human researcher to take full benefit of such an amount of unfiltered data. Moreover, there is a notable and unexplained distinction in the number of results returned. Elliott and Quinn (2013) also refer to the fact that growing amount of available citations is leading to information overload and quote *R v Erskine*, where the court highlighted the need to only cite the cases that established the principle of law, while authorities that were only used to illustrate it should be avoided. It is important to note that information overload may also have a detrimental effect on the quality of analysis offered by humans that read the reports and collate case treatment summaries. Despite the considerable progress made in information management and retrieval, it is believed that compilation of electronic citators is still dependent on human analysis of the case law. Publishers still employ specifically trained humans to provide case analysis necessary for the case treatment tables (Geist 2009). Consequently, it can be expected that under the pressing conditions of information overload the quality of human analysis of case citations can suffer when applied to other types of legal documents, such as case law reports.

Secondly, the complexity of interpretation and variance in conceptualisation can also make processing of citation information more challenging. Ogden (1993a) mentions that Greenleaf, the first publisher of a paper legal citator, found the following:

> ...determining the authority of a case required a strong grasp of precedent and legal analysis, not to mention the stamina required to read all the cases.

The situation has not significantly changed since the times of the first legal citators as Elliott and Quinn (2013) also criticise modern judgments for being often very long, hard to read and containing ratios that are "buried in the sea of irrelevant information". Marmor (2005) observes that overall there is not always sufficient agreement between the legal professionals regarding the case law:

> ...appellate court decisions are rife with disagreements between the judges on what the law is.

The observations made by Marmor (2005) are in agreement with the conclusion recently reached by Greenawalt (2012):

> ...competing conceptualizations may affect how judges and legal scholars think about what actually is, and should be, going on.

An additional aspect is that searching for legal principles supported by a case may be complicated as there has never been a definitive guide on how to correctly interpret a case. The nature and definition of *ratio* has also been a subject of debates, though this will not be explored in detail in this paper. But, as one example, Branting (1994b) claimed that a number of statements satisfying the ratio test can be found in case law reports. As a part of this study, the term ratio will be avoided due to the complexities associated with its definition and identification. Instead, the law that is supported by a certain case will be defined as legal principles. Existing interpretations of case law provided by courts as citations are binding and can be used to illustrate the legal principles that are supported by the cited case and material facts that are necessary for the principles to be invoked. A full manual study aimed at detailed analysis and systematisation of common law principles may, however, be time consuming and not always possible with the available tools.

The discussion above sets the context and motivation for the current study. In the next section, we turn to related work.

3 Related work

This research aims to apply machine learning methodology in order to automatically identify legal principles and facts associated with case citations. A significant amount of work has been done in the area of citation analysis in scientific literature, while only a relatively smaller amount of work has been done that focuses on studying case law citations. Most existing studies on case law citations aim to identify case treatment—the relationship between citing and cited cases (e.g. *distinguished*, *explained*, and others)—or analyse citations from the point of view of network analysis, but don't focus on fine-grained analysis of the cited information. There are few reported works that specifically aim to apply machine learning methodology to identify legal principles and facts of the cited cases in case decisions. In the following subsections, we discuss related work on citation analysis along with relevant literature on legal argumentation.

3.1 Citation analysis

The first attempts to systematise citation information were done in the field of common law by the developers of legal citators, starting with Frank Shepard in 1873, who relied on human expertise to provide discourse-aware summaries of case law citations. More recently, citation information is presented as in LexisNexis Shepard's Citations Service.

Despite lawyers being the pioneers of citation analysis (Ogden 1993b), the research on citation analysis in common law has not been developing as fast as citation analysis in the domain of scientific reports. Garfield (1955) is often cited as one of the pioneers and key contributors towards citation analysis in science.

Garfield was inspired by the Shepard's citations and argued that similar methodologies can be useful for summarisation of scientific citations (Garfield 1955). Garfield employed a bibliographic approach to create ICI Citation Indexes, and the data from citation indexes was later used for a number of bibliometric studies that "extract, aggregate and analyse quantitative aspects of bibliographic information" (Moed 2005). He believed that citation analysis could be used for evaluation of scientific performance, for example, in calculation of journal ranks based on citation frequency and impact. As noted by Moed (2005), quantitative data from bibliometric studies is widely used to assess the performance of individual scholars, scientific journals, research institutions and "general, structural aspects of the scholarly system" (e.g. measuring trends in national publication output). Moed (2005) also concluded that ICI citation indexes do not "capture motives of individuals, but their consequences at an aggregate level" and argued for further development of qualitative citation based indicators, thus abandoning the principle underlying most citation analyses that "all citations are equal". Qualitative approaches in citation analysis take into account the intentions of the person who was providing the citation. They aim to capture citation qualities that are overlooked by quantitative methodologies, for example, such as polarity and sentiment. A scientific article may be frequently cited, but it can be due to criticisms or mere acknowledgements, which distinguishes it from an article introducing an approach that is widely accepted and utilised. Several researchers can be mentioned in respect of qualitative citation based indicators in science (Moravcsik and Murugesan 1975; Swales 1986; Cano 1989; Teufel et al. 2006; Athar and Teufel 2012). Cronin (1982) conducted a research of citation behaviours and noted that at the time there was not a universal approach in citation studies. Application of qualitative citation based indicators often relies on linguistic discourse markers to generate conclusions about citations and citing behaviours. For example, citations can be classified according to sentiment polarities: confirmative or negative (Moravcsik and Murugesan 1975); positive, neutral or weak (Teufel et al. 2006). Aspects of qualitative analysis of citations is relevant to our approach.

Recently there has been more interest toward citation studies in law, where there appear to be two major directions: applying network analysis to citations (Zhang and Koppaka 2007; Leicht et al. 2007; Winkels et al. 2011; Lupu and Voeten 2012; van Opijnen 2012; Neale 2013) and classification systems allowing one to estimate the "treatment" status of the cited case (Jackson et al. 2003; Galgani et al. 2015).

Zhang and Koppaka (2007) developed a Semantics-Based Legal Citation Network (see Zhang et al. (2014) for an overview of related work), a tool that extracts and summarises citation information into a network, allowing the users to "easily navigate in the citation networks and study how citations are interrelated and how legal issues have evolved in the past." The researchers note that different parts of a case can be cited. Studying the reasons for citation can provide valuable information for a legal researcher. Their approach relied on RFC (reason for citing), a patented technology that allows extracting reasons of why the case has been cited. RFC performance was summarised in the patent (Humphrey et al. 2005), which explored a methodology of "identifying sentences near a document citation (such as a court case citation) that suggest the reason(s) for citing (RFC)". The task of

identifying RFC may be somewhat similar to the task that is undertaken as a part of this project due to the fact that information contained in principles and facts of cited cases can be used as a part of estimating reasons for citing. However, the methodology of Humphrey et al. (2005) is largely based on an analysis of word frequencies; there is no machine learning and no evaluation is presented. This contrasts with our approach to identifying the relevant statements with reference to particular features and associating them to a specific citation in a decision in situ; our approach applies machine learning and is evaluated.

History Assistant Jackson et al. (2003) was designed to automatically infer direct and indirect treatment history from case reports. Direct treatment history covered historically related cases, such as appeals etc. Indirect treatment history dealt with the cited cases within a document in order to establish how the cited case has been treated. It relied on the classification methodology of Shepard's citations that combines the knowledge about sentiment and aims of legal communication with heuristic information about court hierarchy. It includes such classes as applied, overruled and distinguished. History Assistant was expected to be an aid for editorial work rather than replace the effort of the editors. The program consisted of a set of natural language modules and a prior case retrieval module. Natural language processing relied on machine learning methodology and employed statistical methods over annotated corpus. The tools and corpora are unavailable for replication. While Shepardisation is intrinsically interesting and difficult task, we found little overt textual evidence to reliably facilitate machine learning.

Galgani et al. (2015) created LEXA—a system that relied on RDR (Ripple Down Rules) approach to identify citations within the "distinguished" class. This category is generally best linguistically signaled and is therefore suitable for achieving high precision and recall. The key idea underpinning RDR was that the "domain expert monitors the system and whenever it performs incorrectly he signals the error and provides as a correction a rule based on the case which generated the error, which is added to the knowledge base" (Galgani et al. 2015). The approach employed annotators to create an initial set of rules leaving the end users to refine and further expand the set. The authors claimed that "the user can at any stage create new annotations and use them in creating rules" which may put a more significant reliance on the user input than an end user may be equipped or expecting to provide. LEXA employed 78 rules that recognized "distinguished" citations with a precision of 70% and recall of 48.6% on the cleaned test set, which is significantly lower than the results reported by Jackson et al. (2003) for the same category: precision (94%) and recall (90%). The difference in results suggests that a complex fine-grained analysis used by Jackson et al. (2003) that included machine-learning for language processing may help achieve better classification outcomes. Our study does not consider classifications of citations.

In Grabmair et al. (2015), the primary objective is to rank documents for Information Retrieval using the Lucene engine. As part of their work, they created a gold standard wherein statements are annotated as legal principles or evidence along with annotations for aspects of subsentential structure; such sentences are treated preferentially when evaluating document relevance to a search query. Some success is reported in automatically annotating the statements using a mix of n-gram

features and entity types, identified through manually coded UIMA rules. Subsequent work in Bansal et al. (2016) provide additional features and report some improvement in performance, particularly in ranking. The work is carried out on a domain specific set of cases (US Special Masters Vaccine Injury Decisions), which may not generalise; our work is domain neutral. The features (n-gram and entity types) used in Grabmair et al. (2015) are not as clearly tied to statement annotations as in our approach. Our approach does not consider ranking or high level properties of decisions, but tying specific statements of legal principles and evidence to citations within decisions; our results are promising.

3.2 Argument extraction

There have been a variety of attempts aimed at automated extraction of argumentation structure of text and its constituents. Clearly the identification of statements of legal principle and evidence tied to a citation in a decision is relevant to the presentation of the overall legal argument. However, argumentation structure is a larger scale matter within which we can locate our specific study.

The methodologies employed by such studies often rely on extraction and further analysis of linguistic information that is available within the text. One of the relatively recent successful examples of argumentation extraction methodology can be argumentation zoning. This approach is based on the assumption that the argumentation structure can be presented as a combination of rhetorical zones that are used to group the statements according to their rhetorical role. This approach was initially used for scientific reports (Teufel et al. 1999, 2009). Hachey and Grover (2006) used argumentation zoning to create summaries for common law reports. Note that argumentative zoning, both for science (Teufel et al. 1999) and for law (Hachey and Grover 2006), is very much aimed at capturing the argumentation used by the author of that paper or judgement. All sentences summarising previous papers or judgements are annotated as a single category (OTHER (Teufel et al. 1999) or PROCEEDINGS (Hachey and Grover 2006). Our work differs in that we aim to mine descriptions of previous judgements for facts and principles. With respect to the methodology used for automatic classification of sentences, both these studies rely on manually constructed linguistic knowledge bases (e.g. a categorisation of cue phrases used in science and regular expressions for matching them (Teufel et al. 1999), or specially developed Named Entity Tools to identify judges, appellants, respondents, etc. (Hachey and Grover 2006), and report acceptable results for most of the categories, with some categories performing better than others. We do not rely on such manually developed resources, and restrict ourselves to linguistic features derived from the sentences themselves. We do, however, follow these studies closely in the methodology used to manually annotate our corpus and report agreement between annotators.

An approach similar to argumentation zoning was taken by Farzindar and Lapalme (2004) to develop a scheme for identification of argument structure of Canadian case law and Kuhn (2010) to analyse the structure of German court decisions. A methodology relying on manual annotation of discourse structures and in that respect similar to argumentation zoning was used by Wyner (2010) to detect

case elements such as case citation, cases cited, precedential relationships, names of parties, judges, attorneys, court sort, roles of parties (i.e. plaintiff or defendant), attorneys, and final decision. Whilst the methodology developed does not aim to fully reconstruct argumentation structure, the information obtained during the study can be used as a part of a wider application.

Wyner et al. (2010) conducted a study aimed at identification of argumentation parts with the use of context-free grammars. Similar to Jackson et al. (2003) the study reports the following difficulties with identifying argumentation structures in legal texts: "(a) the detection of intermediate conclusions, especially the ones without rhetorical markers, as more than 20% of the conclusions are classified as premises of a higher layer conclusion; (b) the ambiguity between argument structures." The results reported are as follows: premises—59% precision, 70% recall; conclusions—61% precision, 75% recall; non-argumentative information—89% precision, 80% recall.

The methodology of applying statistical tools over annotated corpus was employed by Moens et al. (2007) to automatically detect sentences that are a part of the legal argument. The study achieved 68% accuracy for legal texts. Ashley and Walker (2013) aimed to extract "argumentation-relevant information automatically from a corpus of legal decision documents" and "build new arguments using that information".

A related, important distinction that should be made with regard to legal argumentation is the idea that the cited legal principles can be classed as *ratio* or *obiter*. As defined by Raz (2002): "ratio decidendi can be understood as those statements of law which are based on the facts as found and upon which the decision is based." Statements that are usually included into obiter class are dissenting statements and statements that are "based upon either nonexistent or immaterial facts of the case" (Raz 2002). From the point of view of law the main difference between *ratio* and *obiter* is that the former is binding, while the latter only possesses persuasive powers. Branting (1994a) tried to automatically identify and extract *ratio*. Plug (2000) tried to identify *obiter* statements. However, the distinctions between *ratio* or *obiter* will not be used as a part of this work.

4 Manual annotation study

The manual annotation study focused on annotating the gold standard corpus and evaluating the annotation methodology to confirm that the defined categories could be reliable distinguished by human annotators based on written guidelines. This gold standard corpus is then used for the machine annotation study in Sect. 5. Two annotators were used for the purposes of the manual annotation agreement study: Annotator 1 and Annotator 2. Annotator 1 has legal training and Annotator 2 does not. All manual annotation was performed in GATE[3], a widely used text analysis tool.

[3] GATE 8.0: https://gate.ac.uk.

4.1 Method

The corpus for the gold standard was compiled from 50 common law reports that had been taken from the British and Irish Legal Institute (BAILII) website in RTF format. Most reports used for this study only provided the leading opinion and were narrated by the court in the form of monologue speech. The topics that were covered by the selected reports were related to civil matters, mainly issues covered by contract, trust and property law.[4] The length and structure of reports varied, which was most often defined by the complexity of the matter: longer and more complicated cases often had more sections. As reported by GATE Sentence Splitter (GATE 8.0.), the full corpus contained 1211012 tokens (or words) and 22617 sentences which included headings and other units that didn't form full sentences from grammatical point of view. Most reports had a section on the top introducing the court, the parties, legal representatives, case number etc. It was often the case that the legal situation was presented in the introduction and that the legal analysis was in the middle of the report. However, the reports did not follow a universal format. Conclusions were often short and situated at the end of the report. Case law citations are used to support legal argumentation and are therefore referred to as a part of legal analysis. For that reason they were rarely found in introduction or conclusion.

Annotator 1 created annotation guidelines (high level task definition, descriptions and examples for each category, and analyses of a few difficult cases) in several iterations and trained Annotator 2 on their use. The annotators were expected to identify sentences that contained the legal *principles* and *facts* of the cited cases, based on the written guidelines. Sentences associated with cited cases that are neither *principles* or *facts* are annotated as *neutral*. The key points of the guidelines, which were 7 pages long, are summarised below.

The task of annotation focused on the identification of cited information within annotation areas that were defined as paragraphs having at least one citation. Citation instances had been manually annotated prior to the study. The proportion of the sentences in the gold standard corpus that were annotated within the annotation areas corresponds to around 12% (or 2659 sentences). The control corpus used for the human agreement study contained 301486 tokens, 5716 sentences (25% of the gold standard corpus) and 241 references to citation names (29% of the gold standard corpus). The proportion of sentences in the control corpus that were situated within the annotation areas was 14% (821 sentences in total).

Given the discussion of the complexity of jurisprudential views of legal principles, we have taken an operationalised view, based on the analysis of a legal scholar and key linguistic indicators. A nine page annotation manual was produced to define each category and provide examples and counter examples for difficult cases. The key points are summarised here.

A *legal principle* is defined as any statement which is used, along with facts, to reach a conclusion. Linguistically, a legal principle can, for instance, be indicated by deontic modality, e.g. expressions of *must* for obligation, *must not* for

[4] The corpus is available upon request.

prohibition, or *may* for permission, which contrast with epistemic modalities for necessity and possibility. For example:

> As a matter of principle no order *should* be made in civil or family proceedings without notice to the other side unless there is a very good reason for departing from the general rule that notice *must* be given (Gorbunova v Berezovsky (aka Platon Elenin) & Ors, 2013).

Legal principles can be qualified, e.g. with conditions that may limit the application of rule. It is also possible that legal principles are "active" in reasoning, yet inferred from the text, in which case, they cannot be annotated or used for further text processing.

In contrast to legal principles, there are *facts*, which are statements bearing on what uncontroversially exists, occurred, or is a piece of information. For our purposes, only sentences that refer to events which occur outside the court hearing are annotated; this excludes procedural facts. For example:

> Miss Lange was not a party to the 1965 Transfer or the 1968 Deed and she covenanted only with Mrs de Froberville (and not with Brigadier Radford) to comply with the covenants in those instruments in so far as they were still subsisting and capable of taking effect (89 Holland Park (Management) Ltd & Ors v Hicks, 2013).

Linguistically, facts present themselves with non-modal expressions and denoting expressions, e.g. uses language which is specific, actual, and definite.

Following a period of training, a set of 10 reports were randomly selected (all previously unseen by the annotators) as the aforementioned control corpus for the inter-annotator and intra-annotation agreement studies reported here. The process in short was to:

1. Use the pre-annotated citation instances to identify annotation areas—i.e. paragraphs that contain at least one citation name. Direct quotes and lists were treated as a part of the same paragraph.
2. Label each sentence in each annotation area as one of *fact*, *principle* or *neither*, following the annotation guidelines.

4.2 Results

Table 1 shows the distribution of categories in the evaluation set of 10 reports. It shows that Annotator 2, who does not have legal training, is more conservative in identifying facts and inferences than Annotator 1, who has had legal training.

The results of the inter-annotator agreement study are as follows: $\kappa = 0.65$[5] (% Agreement = 83.7). The intra-annotator agreement study showed that Annotator

[5] κ, the predominant agreement measure used in natural language processing research (Carletta 1996), corrects raw agreement $P(A)$ for agreement by chance $P(E)$: $\kappa = \frac{P(A)-P(E)}{1-P(E)}$.

Table 1 Distribution of categories

	Annotator 1 (original annotation)	Annotator 2 (inter-annotator study)	Annotator 1 (intra-annotator study)
Principles	266 (32%)	211 (26%)	258 (31%)
Facts	56 (7%)	20 (2%)	54 (7%)
Neither	499 (61%)	590 (72%)	509 (62%)

1 (when annotating the same set of 10 reports three months apart in time) was extremely consistent: $\kappa = 0.95$ (% Agreement = 97.3).

Annotator 1 proceeded to create a gold corpus of 50 reports which was used for training a machine classifier, as described next.

5 Automated annotation study

The methodology used for machine annotation employed classification of the annotation units with a Naive Bayesian Multinomial Classifier based on a set of selected features described below.

5.1 Features for classification

The task of features selection focused on identifying the features that can help in classifying sentences. The following features were selected for extraction from the dataset:

- Part of speech tags.
- Unigrams.
- Dependency pairs.
- Length of the sentence.
- Position in the text.
- Cit—a feature which indicates whether there is a citation instance in the sentence.

Unigrams are widely used in text classification tasks. The performance of classifiers relying on bag-of-words approach can however be impeded by the assumption that words are independent; i.e., grammatical relations are not significant. To address this limitation researchers often complement unigrams with features that can capture dependencies between words. Dependency pairs derived using the Stanford Parser (Marneffe et al. 2006) were used to complement unigrams, creating word pairs that are grammatically linked rather than simply collocated like n-grams. Dependency features have previously been shown to be difficult to beat for a variety of text classifications tasks such as sentiment analysis (Joshi and Penstein-Rosé 2009) and stance classification (Hasan and Ng 2014; Mandya et al. 2016).

Part of speech tags were selected as a feature for a number of reasons. Firstly, it was expected that modal verbs and verb tense may help to classify the annotation

units. Sentences that introduce facts are most often presented in the Past tense. For example:

> The contract **contained** a general condition that in relation to any financial or other conditions either party could at any time before the condition **was fulfilled** or **waived** avoid the contract by giving notice.

Secondly, both epistemic and deontic modal qualifiers that use modal verbs are common in sentences containing legal principles, for example:

> "It is a question which **must** depend on the circumstances of each case, and mainly on two circumstances, as indicating the intention, viz., the degree of annexation and the object of the annexation" (Cardigan v Moore & Anor, 2012).

> "As a matter of principle no order **should** be made in civil or family proceedings without notice to the other side unless there is a very good reason for departing from the general rule that notice must be given" (Gorbunova v Berezovsky (aka Platon Elenin) & Ors, 2013).

In addition, we used three other features that captured the length of the sentence (number of words), its position in the text (on a scale of 0–1) and whether or not there is a citation in the sentence (boolean).

We used NLTK (Bird 2006) to extract part of speech tags and Stanford CoreNLP (Manning et al. 2014) to extract grammatical relations or dependencies. The other features were derived by means of a python script.

5.2 Machine learning framework

Our machine learning experiments were conducted using Weka (Hall et al. 2009), a collection of machine learning algorithms for data mining tasks. Given the limited amount of labeled data available, there was no developmental stage employed. We instead used linguistic features that we expected to be useful, relied on automatic feature selection to prune the feature set, ran a single machine learning algorithm with default settings and report results using a cross-validation methodology, as detailed below:

1. Feature counts were normalised by tf and idf.
2. Attribute selection (InfoGainAttributeEval in combination with Ranker (threshold = 0) search method) was performed over the entire dataset.
3. The Naive Bayes Multinomial classifier was used for the classification task. This has been widely used in text classification tasks (Teufel et al. 2006; Mitchell 1997), and its performance is often comparable to more sophisticated learning methods (Schneider 2005).
4. Results are reported for tenfold cross-validation. The 2659 sentences in the dataset were randomly partitioned into 10 subsamples. In each fold one of the subsamples was used for testing after training on the remaining 9 subsamples. Results are reported over the 10 testing subsamples, which constitute the entire dataset.

5.3 Results

Tables 2 and 3 report the classification performance of the Naive Bayes Multinomial classifier from the Weka toolkit (Hall et al. 2009). Feature selection reduced the number of features from 51576 to 887; we report more on the selected features later in this section.

The accuracy of the classifier is slightly better than that of the Annotator 2 (as reported in Sect. 4.2), who had no legal training in the manual study. The classifier achieves high precision and recall for each of the three categories, despite the unbalanced nature of the corpus (60% neutral, 30% principles and 10% facts). The confusion matrix in Table 3 shows that *facts* and *principles* are distinguished easily from each other. The majority of classification errors involve confusion with the *neither* category. These results suggest that to the extent such annotations can be carried out based on linguistic principles alone, automated annotation can be performed to the same standard as manual annotation.

Table 4 shows the top 100 features for this classification task. These are mostly either part-of-speech tags, unigrams such as 'is', 'be', 'was', 'must', 'may', 'will' and 'can' that indicate tense or modality, unigrams such as 'a', 'an' and 'the' that

Table 2 Per category and aggregated statistics for automatic classifier

	Precision	Recall	F-measure
Principles	0.823	0.797	0.810
Facts	0.822	0.815	0.818
Neither	0.877	0.892	0.884
Number of Sentences	26.59		
Accuracy	085		
κ	0.72		

Table 3 Confusion matrix

Classified as ⟶	Principles	Facts	Neither
Principles	646	5	160
Facts	4	198	41
Neither	135	38	1432

Table 4 Top 100 features by information gain

Part-of-speech tags: VBZ, NN, JJ, VB, MD, DT, IN, CC, NNP, VBN, COMM, RB, WRB, SEMM, FS, NNS, TO, QUOT, WDT, VBG, WP, POS, VBP

Unigrams: is, a, the, or, be, Mr, was, must, of, had, may, has, see, 0, it, 100, where, I, were, other, are, if, will, to, concerned, general, person, 300, an, Mrs, and, judgment, party, that, planning, principle, letter, company, one, If, circumstances, which, per, money, whether, Hawk, always, submissions, not, Charles, Arista, court, jurisdiction, forum, can, pictures, Akzo, Miss, ordinary, fund, man, S1, contained, 000, wholesalers, this

Dependency pairs: case-that, is-there, concern-was, concern-case, case-was, parties-the, condition-a, court-the, is-if, letter-the

Table 5 Performance of each type of feature

	Majority class	Part-of-speech	Unigrams	Dependencies	All
Accuracy	0.60	0.63	0.77	0.81	0.85
κ	0.00	0.18	0.58	0.63	0.72

indicate definiteness, unigrams such as 'if', 'whether' and 'where' that can be used in stating conditions, as well as unigram and dependency pair features involving generic legal words such 'principle', 'judgment', 'concerned', 'party' 'court', 'jurisdiction', 'case', 'condition' etc. Only a small number of the features represent noise from overfitting the training data, including proper names such as 'Charles' and 'Akso' and numbers such as '0', '100', '300'. Using only these 100 features already achieves a reasonably high accuracy of 0.72 ($\kappa = 0.48$), comfortably outperforming the majority class baseline (accuracy = 0.60, $\kappa = 0.00$).

In the learnt Bayesian model, part of speech features such as 'MD' (modals), 'VBZ' and 'VB' (present tense), 'WRB' (WH-adverbs), as well as unigrams such as 'a' and 'an' (indefinites), 'is', 'be' and 'has' (present tense), 'may','must', 'will' and 'can' (modals), 'if', 'whether', 'or', 'unless' and 'where' (conjunctions) have higher probability for the the *principles* class, as do general purpose nouns such as 'person' and 'party'. This is intuitive as principles are often stated in present tense, scoped with a modal verb, presented in general terms (thus using indefinites or general nouns such as 'person' or 'party') and also typically relate more than one clause using conjunctions. On the other hand, the main features that predict *facts* are the use of past tense ('VBD', 'VBN','was', 'were', 'had', 'concerned', etc.) and proper names ('NNP', 'Mr', 'Mrs', etc). The principle features that select for the *neither* class are the use of the first person ('I', 'my' and 'judgment-my'), other references that indicate the sentence is about the current case rather than a cited one ('this', 'case-this', etc), words indicating that the judge is summarising or quoting ('says', 'says-he', 'summarised') and the use of adjectives and non-WH adverbs ('JJ' and 'RB'), which might indicate an opinion being expressed.

Finally, we evaluated the performance of the classifier on each type of feature (part of speech, unigram and dependency) separately, as reported in Table 5 along with the majority class baseline that labels all sentences as *neutral*. While using only part of speech tags achieves 63% accuracy, dependency features by themselves achieve 81% accuracy. The combination of all three feature types results in the best results (85% accuracy). All feature sets outperform the baseline.

5.4 Error analysis

A simple visual inspection of confusion output was performed and some hypotheses regarding the causes of confusion were made. In the gold standard corpus a variety of sentences containing facts have been annotated, which included sentences whose main purpose is other than introducing facts. In real life scenarios courts don't always provide a detailed description of facts, but instead embed facts within legal reasoning. For this reason, sentences that contain the information about facts follow

a wide variety of grammatical patterns. In combination with a relatively small amount of available instances such variety must have had a negative impact on the classification outcomes. For example, the machine annotator failed to identify the following statement as containing a fact:

> "In Antec International Limited v Biosafety USA Inc [2006] EWHC 47 Mrs Justice Gloster was dealing with an application to set aside an order giving leave to serve abroad in a contractual claim where the contract contained a non-exclusive jurisdiction clause." (Abela & Ors v Baadarani & Anor, 2011)

Visual inspection of instances suggests that confusion between fact and principle though rare overall may be typical in sentences whose aim is not to introduce facts and where factual information is used as a part of reasoning. Such sentences often only contain a short clause containing information about facts, so that in a small dataset, statistical weights associated with the rest of the sentence may outweigh those associated with the clause. For example:

> "The fact that the parties have freely negotiated a contract providing for the non-exclusive jurisdiction of the English courts and English law creates a strong prima facie case that the English jurisdiction is the correct one" (Abela & Ors v Baadarani & Anor, 2011).

The main cause of error for the automatic annotation of principles was that the gold standard only annotated principles from cited cases, but often these were linguistically indistinguishable (in our machine learning approach) from discussions of principles by the current judge; i.e principles expressed by the current judge should have been annotated as neither, but were frequently annotated as principles.

Better results may be achieved if the annotation guidelines are redefined to be more specific about what constitutes a fact or principle, for instance, the fact class could be limited only to the sentences whose aim is to introduce facts. Introducing further features to determine the provenance of principles could help with the confusions between principles and neutral.

6 Conclusions and future work

An overall analysis suggests that the machine annotation experiment has returned good classification results with the Naive Bayesian Multinomial classifier identifying 85% of instances correctly and achieving Kappa equal 0.72. Good combinations of precision and recall have been achieved for all categories (rounding): 82% precision and 80% recall (principles), 82% precision and 81% recall (facts), and 87% precision and 89% recall (neither). Such positive results suggest that the methodology employed as a part of this experiment can provide a suitable basis for further work.

Lawyers use case law citation to refer to existing legal principles. This practice stems from the doctrine of *stare decisis* that prescribes for the cases that are similar on facts to be treated in a similar way. There is no established formal methodology prescribing how legal principles should be extracted from the case or which facts

should be treated as material for making the decision. Citations are a valuable source of existing interpretations of case law which can be used to illustrate the legal principles that are supported by the cited case and material facts that are necessary for the principles to be invoked. This information is important for legal researchers, because it allows identifying a pool of relevant case law that can be used to build the argument. Automated analysis of legal principles and facts within cited cases allows identifying the key information about the cited case which can be used for many purposes, including creation of detailed case treatment summaries, improvement of search and retrieval methodology for the case law and many others.

This work demonstrates the feasibility of automatic identification of legal principles and facts that are associated with a case citation. This functionality could, for example, allow a legal practitioner to not only search, say in Google, for citations mentioned in a case, but also the associated legal principles and facts, providing deep access to and insight into the development of the law. It would also offer the opportunity to access the law directly rather than via the edited and structured materials made available by legal service providers. Finally, we have only addressed accessing cited legal principles and facts, which is distinct from ranking and relating precedents, i.e. Shepardisation. In future work, the source material annotated here could be used to investigate the automation of Shepardisation as well.

References

Ashley KD, Walker VR (2013) Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In: Proceedings of the fourteenth international conference on artificial intelligence and law, ACM, pp 176–180

Athar A, Teufel S (2012) Context-enhanced citation sentiment detection. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies, Association for Computational Linguistics, pp 597–601

Bansal A, Bu Z, Mishra B, Wang S, Ashley K, Grabmair M (2016) Document ranking with citation information and oversampling sentence classification in the luima framework. In: Legal knowledge and information systems: JURIX 2016: the twenty-ninth annual conference, IOS Press, pp 33–42

Bird S (2006) Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on interactive presentation sessions, Association for Computational Linguistics, pp 69–72

Branting K (1994a) Four challenges for a computational model of legal precedent. THINK (J Inst Lang Technol Artif Intell) 3:62–69

Branting L (1994b) A computational model of ratio decidendi. Artif Intell Law 2:1–31

Cano V (1989) Citation behavior: classification, utility, and location. J Am Soc Inf Sci 40(4):284

Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 22(2):249–254

Cooper BD (1982) Anglo-american legal citation: historical development and library implications. Law Libr J 3:3–33

Cronin B (1982) Norms and functions in citation: the view of journal editors and referees in psychology. Soc Sci Inf Stud 2(2):65–77

De Marneffe MC, MacCartney B, Manning CD et al (2006) Generating typed dependency parses from phrase structure parses. Proc LREC 6:449–454
Elliott C, Quinn F (2013) English legal system, 14th edn. Pearson Education Inc, Karnataka
Farzindar A, Lapalme G (2004) Legal text summarization by exploration of the thematic structures and argumentative roles. In: Text summarization branches out workshop held in conjunction with ACL, pp 27–34
Galgani F, Compton P, Hoffmann A (2015) Lexa: building knowledge bases for automatic legal citation classification. Expert Syst Appl 42(17):6391–6407
Garfield E (1955) Citation indexes for science. Science 122:108–111
Geist A (2009) Using citation analysis techniques for computer-assisted legal research in continental jurisdictions. PhD thesis, University of Edinburgh
Gerken J (2016) The invention of legal research. William S. Hein and Co., Getzville
Grabmair M, Ashley KD, Chen R, Sureshkumar P, Wang C, Nyberg E, Walker VR (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 and law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015, pp 69–78
Greenawalt K (2012) Statutory and common law interpretation. Oxford University Press, Oxford
Greenawalt K (2013) Interpretation and judgment. Yale J Law Humanit 9(2):5
Hachey B, Grover C (2006) Extractive summarisation of legal texts. Artif Intell Law 14(4):305–345
Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD Explor Newsl 11(1):10–18
Hasan KS, Ng V (2014) Why are you taking this stance? Identifying and classifying reasons in ideological debates. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp 751–762. http://www.aclweb.org/anthology/D14-1083
Humphrey TL, Lu XA, Parhizgar A, Ahmed S, Wiltshire Jr JS, Morelock JT, Harmon JP, Collias SG, Zhang P (2005) Automated system and method for generating reasons that a court case is cited. US Patent 6,856,988
Jackson P, Al-Kofahi K, Tyrrell A, Vachher A (2003) Information extraction from case law and retrieval of prior cases. Artif Intell 150(1):239–290
Joshi M, Penstein-Rosé C (2009) Generalizing dependency features for opinion mining. In: Proceedings of the ACL-IJCNLP 2009 conference short papers, Association for Computational Linguistics, pp 313–316
Kuhn F (2010) A description language for content zones of German court decisions. In: Proceedings of the LREC 2010 workshop on the semantic processing of legal texts, pp 1–7
Leicht EA, Clarkson G, Shedden K, Newman ME (2007) Large-scale structure of time evolving citation networks. Eur Phys J B 59(1):75–83
Lupu Y, Voeten E (2012) Precedent in international courts: a network analysis of case citations by the european court of human rights. Br J Politic Sci 42(02):413–439
Mandya A, Siddharthan A, Wyner A (2016) Scrutable feature sets for stance classification. In: Proceedings of the 3rd workshop on argument mining, ACL 2016, Association for Computational Linguistics, Berlin
Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D (2014) The Stanford Cornell natural language processing toolkit. In: ACL (system demonstrations), pp 55–60
Marmor A (2005) Interpretation and legal theory. Hart Publishing, Oxford
Mart S (2013) The case for curation: the relevance of digest and citator results in westlaw and lexis. Leg Ref Serv Q 32(1–2):13–53
Mitchell TM (1997) Does machine learning really work? AI Mag 18(3):11
Moed HF (2005) Citation analysis of scientific journals and journal impact measures. Curr Sci 89(12):1990–1996
Moens MF, Boiy E, Palau RM, Reed C (2007) Automatic detection of arguments in legal texts. In: Proceedings of the 11th international conference on Artificial intelligence and law. ACM, pp 225–230
Moravcsik MJ, Murugesan P (1975) Some results on the function and quality of citations. Soc Stud Sci 5:88–91
Neale T (2013) Citation analysis of canadian case law. J Open Access Law 1:1
Ogden P (1993a) Mastering the lawless science of our law: a story of legal citation indexes. Law Libr J 85(1):1–48

Ogden P (1993b) Mastering the lawless science of our law: a story of legal citation indexes. Law Libr J 85:1

Plug J (2000) Indicators of obiter dicta. a pragma-dialectical analysis of textual clues for the reconstruction of legal argumentation. Artif Intell Law 8(2–3):189–203

Raz M (2002) Inside precedents: the ratio decidendi and the obiter dicta. Common Law Rev 3:21

Schneider KM (2005) Techniques for improving the performance of naive bayes for text classification. In: International conference on intelligent text processing and computational linguistics. Springer, pp 682–693

Someren A (2014) Finding a categorization in references from legislation. PhD thesis, University of Amsterdam, Amsterdam, The Netherlands, bachelor thesis

Swales J (1986) Citation analysis and discourse analysis. Appl Linguist 7(1):39–56

Teufel S, Carletta J, Moens M (1999) An annotation scheme for discourse-level argumentation in research articles. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp 110–117

Teufel S, Siddharthan A, Tidhar D (2006) Automatic classification of citation function. In: Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 103–110

Teufel S, Siddharthan A, Batchelor C (2009) Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In: Proceedings of the 2009 conference on empirical methods in natural language processing, vol 3. Association for Computational Linguistics, pp 1493–1502

van Opijnen M (2012) Citation analysis and beyond: in search of indicators measuring case law importance. JURIX 250:95–104

Winkels R, Ruyter JD, Kroese H (2011) Determining authority of dutch case law. Leg Knowl Inf Syst 235:103–112

Wyner AZ (2010) Towards annotating and extracting textual legal case elements. Inf Dirit Spec Issue Leg Ontol Artif Intell Tech 19(1–2):9–18

Wyner A, Mochales-Palau R, Moens MF, Milward D (2010) Approaches to text mining arguments from legal cases. In: Francesconi E, Montemagni S, Peters W, Tiscornia D (eds) Semantic processing of legal texts. Lecture notes in computer science, vol 6036. Springer, Heidelberg

Zhang P, Koppaka L (2007) Semantics-based legal citation network. In: Proceedings of the 11th international conference on artificial intelligence and law, ACM, pp 123–130

Zhang P, Silver H, Wasson M, Steiner D, Sharma S (2014) Knowledge network based on legal issues. In: Winkels R (ed) Proceedings of workshop on network analysis in law (NAilL2013), pp 23–28