Artif Intell Law
https://doi.org/10.1007/s10506-018-9217-1

Automatic semantic edge labeling over legal citation graphs

Ali Sadeghian[1] · Laksshman Sundaram[1,2] · Daisy Zhe Wang[1] · William F. Hamilton[1] · Karl Branting[3] · Craig Pfeifer[3]

Abstract A large number of cross-references to various bodies of text are used in legal texts, each serving a different purpose. It is often necessary for authorities and companies to look into certain types of these citations. Yet, there is a lack of automatic tools to aid in this process. Recently, citation graphs have been used to improve the intelligibility of complex rule frameworks. We propose an algorithm that builds the citation graph from a document and automatically labels each edge according to its purpose. Our method uses the citing text only and thus works only on citations who's purpose can be uniquely identified by their surrounding text. This framework is then applied to the US code. This paper includes defining and evaluating a standard gold set of labels that cover a vast majority of citation types which

This work is partially supported by UF CISE Data Science Research Lab, UF Law School and ICAIR Program.

✉ Ali Sadeghian
asadeghian@ufl.edu

Laksshman Sundaram
lakss@stanford.edu

Daisy Zhe Wang
daisyw@ufl.edu

William F. Hamilton
hamiltonw@law.ufl.edu

Karl Branting
lbranting@mitre.org

Craig Pfeifer
cpfeifer@mitre.org

[1] University of Florida, Gainesville, FL, USA

[2] Stanford University, Stanford, CA, USA

[3] MITRE Corp., McLean, VA, USA

Published online: 01 March 2018

appear in the "US Code" but are still short enough for practical use. We also proposed a novel linear-chain conditional random field model that extracts the features required for labeling the citations from the surrounding text. We then analyzed the effectiveness of different clustering methods such as K-means and support vector machine to automatically label each citation with the corresponding label. Besides this, we talk about the practical difficulties of this task and give a comparison of human accuracy compared to our end-to-end algorithm.

Keywords Legal citation graph · Semantics · Automatic citation analysis · Conditional random fields · Word embeddings · Clustering

1 Introduction

New regulations and laws in the United States are legislated or existing ones are evolved through a complex legal cycle. This system involves numerous organizations, parties, and individuals. Individual legal rules seldom exist in isolation, but instead typically occur as components of broader statutory, regulatory, and common-law frameworks consisting of numerous interconnected rules, regulations, and rulings. The complexity of these frameworks impedes comprehension and compliance by government agencies, businesses, and citizens and makes amending legislation laborious and error-prone for regulatory agencies and legislative drafters.

Citation networks are a promising recent approach to improving the intelligibility of complex rule frameworks. In a citation network, rules are represented by nodes and citations are represented by edges (Neale 2013; Zhang and Koppaka 2007). Citation networks can often permit a complex regulatory framework to be comprehended at a glance. Techniques for automatically representing and displaying citation networks is an active area of research.

Computer assisted and automatic systems have been and are growing rapidly in every field (Mollalo et al. 2015; Sadeghian et al. 2015; Rodrıguez et al. 2016; Amir Sadeghian et al. 2017; Cao et al. 2017; Jain et al. 2017; Sharghi 2017). The legal domain is also no exception to this trend (Harrington 1984; Roitblat et al. 2010; Galgani and Hoffmann 2010; Branting 2017; Prakken 1993). Specially there has been extensive research in designing programs and intelligent software that can address the challenging and expensive task of information extraction from general text. Information extraction is of special importance from a legal perspective since almost all the information in this domain is collected in natural human language. These techniques can be utilized to aid in the automation of creating and displaying meaningful citation networks.

An important aspect of citation-network use is that, generally, only a small subgraph is relevant for any particular application or task. Indeed, visualizations of entire citation networks are generally incomprehensible "hairballs".

The subgraph of a citation network relevant to a particular task depends both on the attributes of the nodes (i.e., rules) and edges (i.e., citations). For example, a subgraph relevant to public health emergencies would include both nodes defining

the powers and duties of agents (e.g., doctors, epidemiologists, coroners) and citations indicating the relative authority of these agents. In general, the portion of a statutory framework relevant to given task consists of the subgraph induced by nodes and edges having a semantic relationship to the task.

While nodes relevant to a given task (e.g., UAV licensing) can typically be found using information-retrieval techniques, such as term-vector or topic similarity, identification of relevant edges, is much less well understood. Various researchers have proposed different taxonomies of edges in citation graphs (Hamdaqa and Hamou-Lhadj 2009; Maxwell et al. 2012; Breaux and Antón 2007), but there is not yet a consensus on the most useful set of edge types. Moreover, there has been little progress in automatically applying semantic labels to citations edges, which is essential for large-scale citation network visualization and analysis tools.

This paper first reviews the related work in Sect. 2. Followed by precisely describing our research problem in Sect. 3 and the proposed automated system to tackle this problem in Sect. 4. In Sect. 5, we describe the dataset used to evaluate our system as well as the proposed gold standard label set used for labeling the citation graph. And finally we conclude the paper in Sect. 6 by a summary of the results and a plan for future research on this study.

2 Related work

There have been various previous research projects addressing the detection, resolution, and labeling of citations in the legal domain. But to the best of our knowledge, there has not been any prior work on a systematic approach to automatically detecting the purpose (i.e., label) of cross references with a detailed semantic label set.

In the closest work to ours, Hamdaqa and Hamou-Lhadj (2009) lay the grounds and propose techniques for analysis of citation networks. One of their key contributions is to review methods of automatically detecting the presence of citation in legal texts. They note that even the simple sounding task of detecting a citation is not easy. Although there have been numerous standards and books devoted to proper citation, in many cases the citation text does not follow the correct format and style thus making it hard for automatic extraction of citations from legal documents. They also propose a categorization schema for citations which groups a citation as either an Assertion or an Amendment. They elaborate this categorization in their second paper (Hamdaqa and Hamou-Lhadj 2011). We will discuss more on this later in this section.

In a more recent work, Adedjouma et al. (2014) study and investigate the natural language patterns used in cross-reference expressions to automatically detect and link a citation to its target. One of their main contributions is in the detection of complicated cross references that are written in natural language. But, unlike us, they do not approach the task of labeling the citations and limit their work on resolving the citation links.

Maxwell et al. (2012) aim to develop a system to help software companies comply with all the regulations. They study the taxonomy of legal-cross references

in the acts related to health-care and financial information systems. They claim to be the first to identify concrete examples of conflicting compliance requirements due to cross-references in legal texts. They analyze different patterns of cross-references that occur in these case studies to obtain seven cross-reference types/labels: *constraint, exception, definition, unrelated, incorrect, general*, and *prioritization* and use grounded theory [the discovery of theory from data (Glaser and Strauss 1967)] to conjecture that this set of labels are generalizable to other legal domains. Their definitions of *constraint*, *exception*, *definition* and *prioritization* are very similar to our "Limitation", "Exception", "Definition", "Delegation of Authority". While their *unrelated* label does not apply to general purpose citation labeling and only points out the cross-references that are not related to laws governing software systems. Although we have a more detailed set of labels, we do not have a label that corresponds to *incorrect* since we do not look at the cited text and thus we are not able to determine if the citation is indeed correctly citing the desired section of the citee. Analyzing the cited text can potentially improve the quality of our labeling but needs collecting additional human labeled data and thus we have left it for future work.

Breaux and Antón (2007) propose "Frame-Based Requirements Analysis Method (FBRAM)". FBRAM is a software which helps generate a context-free markup language. Their system facilitates the creation of a model used to systematically acquire a semi-formal representation of requirements from legal texts. The set of labels used in this work is *Exclusion, Fact, Definition, Permission, Obligation, Refrainment*. Their approach in the paper is quite different from ours. They group/label the text and requirements in the cited text while we are interested in the bigger picture of why the statute is being cited. We must also note that FBRAM is utterly relying on a human analyst and mainly helps only if an analyst manually annotates the whole regulatory document first while we use artificial intelligence and machine learning methods to label cross-references.

In a sequel to their first paper, Hamdaqa and Hamou-Lhadj (2011) explore the relationships between the citing and the cited law. Their work is the closest approach to ours in the sense that they also offer an automated system that classifies each citation based on its semantic role in the context. They give a list of advantages of why one would want to explore the relationships among provisions created through citations. In short: it is useful in understanding the impact of changes in a statute and those depending on it; checking consistencies/conflicts between multiple regulations; eases navigation through statutes and their dependencies. They also propose grouping of each edge into Assertions (Definition, Specification, Compliance) and three subtypes of Amendments. They do so buy analyzing the surrounding text. They claim that using the verb which is directly related to the citation, one can label the citation into one of the two main groups. However they do not talk about the possibility of grouping them to the smaller subgroups nor they give numerical evaluations of the accuracy of their approach. In contrast we label each citation into a more refined set and also provide experimental results.

In "Making sense of legal texts", de Maat et al. (2009) study the structure of references and create a parser that can extract and resolve references that achieves very high accuracies on the Dutch law. They also work on classifying and finding

theories of citations in legal texts. Each citation is classified based on its reason for citing into 5 main categories: *Normative*, *Meta-normative*, *Delegating*, *Life cycle*, and *Informative* references. This categories, in comparison to the categories described in this paper, are on a higher level and each may contain multiple of our classes. For example, *Normative* can contain "Definition" and "Limitation" and "Procedure", *Delegation* is very similar to our label "Authority" and *Life cycle* is similar to our "Amended by" and "Amended to". They label the citations using signal words and patterns and include that in their parser.

Winkels et al. (2014) again show that it is possible to automatically and with high accuracy, extract and resolve citations in Dutch legal texts. They also categorize citations from case law to legislation based on how the case decisions is referring to the legislation. In order to do so, they manually inspected a set of 30 cases and extracted, by hand, patterns to use for clustering the citations. An interesting part of their approach is that they use unsupervised learning methods to cluster the citations. Their method is a promising research direction on this topic because it does not necessarily require a predefined set of labels and large amounts of human annotations.

In another similar work on statutory network analysis, Ashley et al. (2014) have analyzed the interlinked actions in public health system dealing emergencies. They define connections between various organizations involved in public health systems and characterize the way in which they are related based on the statutory laws mandating various actions between them. To help them characterize the nature of the link between the organizations, they define a part of the text as "action", that defines the legal mandate that links these organizations. This is similar to our method of using the "predicate text" to label the citation. The next step there entails the classification of the link between the organizations in the prefixed category with the help of the action defined earlier. The work mostly focuses on network analysis and does not explore opportunities presented by machine learning and NLP, which our methods takes advantage of. Our method learns patterns from the existing links to classify citations where the link is not well established by just the "predicate" tags.

3 Problem statement

A citation graph refers to a graph representation of all the cross-references in a document, which can be to other documents or parts of itself. *Nodes* in a citation graph represent sections of the document and *Edges* represent the citations. Thus the citation network is a directed graph such that there is an edge/link from A to B iff part of statute A is citing a part in statute B.

As described in the previous sections, citations differ in meaning and purpose, and dealing with citations in legal documents is important for swift decision making and clear understanding of the law. We aim to annotate each edge with a semantic label that expresses this meaning or purpose. We propose a system that can label each cross-reference according to a predefined set of labels. For the purposes of this paper we only discuss the US Code and its underlying citation graph, but in general,

our approach can potentially be modified to apply to other similar legal citation graphs.

We want to automatically determine the semantic purpose of a citation by grouping the edges into a set of predefined labels that classify each edge based on its reason for being cited. For example, consider the following clause from the US Code:

> *subsection (a) of this section shall not apply to that portion of the employee's accrued benefit to which the requirements of section 409(h) of title 26 apply*

The cited statute, *section 409(h) of title 26*, imposes a *limitation* to where the obligations of the citing text would apply.

This problem can be broken into three subproblems: (1) detecting the citations in a the text (2) defining a set of labels such that it is comprehensive and at the same time short enough to be practical for use (3) classifying the citations into this set of useful labels. We will first talk about our proposed automated system which can be easily generalized to any corpus of legal text and solves the first and third problems assuming a gold set of labels is already known. We then discuss how our label set is obtained and why we think it can be a good set, also discussing its possible limitations and short comings.

In the next section we will provide a descriptive summary of each part of the overall system.

4 The automated system

As we stated in the previous sections, the main focus of this work is to build a system that can automatically label the edges in a citation graph with a predefined set of labels, each of which represents a possible relationship between the citing provision and the cited. That is, label it with the intended purpose of the citation. The first step towards this goal is to be able to automatically detect the presence and span of each citation in the document. We will next describe our citation extraction method.

4.1 Extracting the citation text

The first step towards building this system is to be able to identify a citation. Cross-references in the legal domain mostly follow standards and predefined templates. The Bluebook (Association 1996) or the newer Citation Manual from US Association of Legal Writing Directors (ALWD) (Pollman and Kane 2000) are among the manuals that contain rules for proper citing of legal texts. But as previously mentioned these rules are not always followed.

To extract the citations from a document (e.g., the US Code), we used a complex regex pattern-matching schema that attempts to locate and identify a variety of known formats for citations. The result is the extraction of a number of known corpora types, which then go through an additional processing schema developed to split each extraction - which can potentially include multiple references to the same

or different corpora, such as "26 USC sections 1, 2, and 3 ..." or "28 USC 121 and 10 CFR"—into individual elements and then re-combine them according to basic citation rules, so that it would produce the following: "26 USC 1", "26 USC 2", "26 USC 3", "28 USC 121" and "10 CFR" as 5 separate references. This step can be easily swapped with any of the state of the art methods of citation extraction such as de Maat et al. (2006) and Tran et al. (2014) which achieve very high accuracies. We don't discuss the performance of our citation extraction since it is not the focus of this study and all the methods used to label the citations after extraction can be applied to the output of any such algorithm.

4.2 Feature extraction

A key idea in this method is our novel feature selection. We find a section of the text related to the citation, the *predicate*, and use this as the main feature in our classification. A predicate is a phrase that contains significant information about the type of citations and helps identify the class of a citation. For the purposes of this work, we define the *predicate* as:

1. The full span of words, that
2. Directly expresses the relationship of the cited provision to something in the current section, and
3. Would make sense if applied to any other provision, i.e., contains nothing specific to the subject matter of the particular section (e.g., funds, exemption), and
4. Expresses as much of the semantics (meaning and purpose) of the relationship as possible without violating 1–3.

For example, in:

...all provisions excluded from this chapter under Section 42 U.S.C 1879 ...

the word *under* is not the full possible span that still satisfies (2)–(4), thus violating criterion (1). The phrase *provisions excluded from this chapter under* includes *provisions*, which is not a relationship but is instead the thing that the citation applies to, violating criteria (2) and (3). However, *excluded from this chapter under* satisfies all 4 criteria.

During the annotation process along with collecting a labeled set of citations we also asked each annotator to tag the span of the corresponding "predicate", which we will talk about in more details in Sect. 5.3.

To automatically extract the *predicate* we designed and trained a linear-chain Conditional Random Field (CRF) on our collected annotated data. A detailed description of CRFs can be found in Lafferty et al. (2001), Sutton and McCallum (2006). The correlated sequential structure of the words in a predicate can be well captured with this type of graphical models, which our experimental results in Sect. 5.4 demonstrate too. To create the features, we manually inspected a sample of the raw text and considered the properties of the predicate. Below are a few of these properties:

First, the predicate is almost always preceding the citation. Second, the predicate usually has a certain part of speech (POS) role. Third, specific words such as *under, defined, amended*, tend to appear more in predicate span. For example:

- *The term "commodity broker" means futures ...or commodity options dealer, as defined in section 348d.* (Preposition-Verb-Preposition),
- *Notwithstanding subsections (a), (b), and (c) of this section and paragraph (2) of this subsection, the court shall disallow any claim for reimbursement or ...* (Preposition),
- *Without regard to the the Richard B. Russell National School Lunch Act (4 U.S.C. 1751 et seq.)* (Preposition-Noun-Preposition),
- *The Secretary shall establish criteria for the grants made under subsection (a) of this section, including criteria relating to...*
- *...are transferred to the Office of Congressional Accessibility Services established under section 2172(a) of this title* (*as amended by section 2251 of this title*) ...

Having these properties in mind and to keep the features as simple as possible , we defined the following set of features for each token:

Exact word features We used the exact lowercase token of each word and its neighboring words (before and after it) as three features for each token. We must note that this and other multi-valued categorical variables were binarized for use in the model.

Is digit feature We used a boolean feature to determine if a token is a digit or not.

Part of speech features Based on the lexical tags produced by NLTK (Bird et al. 2009), each word and its neighboring words were assigned with their corresponding POS tags. In addition to that we used the first two and the last two letters of the tag as additional features for the word and its neighbors. This helps when NLTK produces refined POS, for example NNP and NN might have to be treated the same in detecting the predicates.

Distance to citation features We used 5 boolean features determining the relative position of the word to the target citation. $f_1 = 1$ if the word appears after the citation. $f_2 = 1$ if there are no tokens between the word and the citation. $f_3 = 1$ if there is exactly one token between the word and citation. $f_4 = 1$ if there are more than two words in between. $f_5 = 1$ if there are more than four words in between.

Miscellaneous features Other features used were to determine if the word was at the beginning of a sentence, end of a sentence or if the token is a punctuation.

4.3 Classification

One of our main contributions is the automatic process of labeling citations in a legal citation graph. To achieve this goal we utilize an unsupervised learning algorithm to cluster the citations based on a simple word embedding of the extracted *predicates*.

More precisely we first trained a shallow two layered neural network on a large corpus of English text extracted from wikipedia and fine tuned it by another round

of training on the whole corpus of US Code. This approach is a well known method for representing words as vectors in a high dimensional space of real numbers first introduced by Mikolov et al. (2013b). We then use these vectors as the underlying representation of words in the predicate and cluster them using k-means. Subsequently each citation is labeled based on the cluster representing it. More detailed explanation and experimental evaluations are presented in Sect. 5.

4.4 Complete system

In summary the complete system enables automatic labeling of the citations in a legal document. After the legal document is given to the system input, it detects all the citations present in the document using the methods described in Sect. 4.1. It then automatically extracts what we call the predicate which contains information about the type of the at hand citation, this step was described in Sects. 4.2 and 5.4. In the next step it utilizes machine learning techniques described in Sects. 4.3 and 5.5 to assign to the citation, an appropriate label. The final labeled graph is then

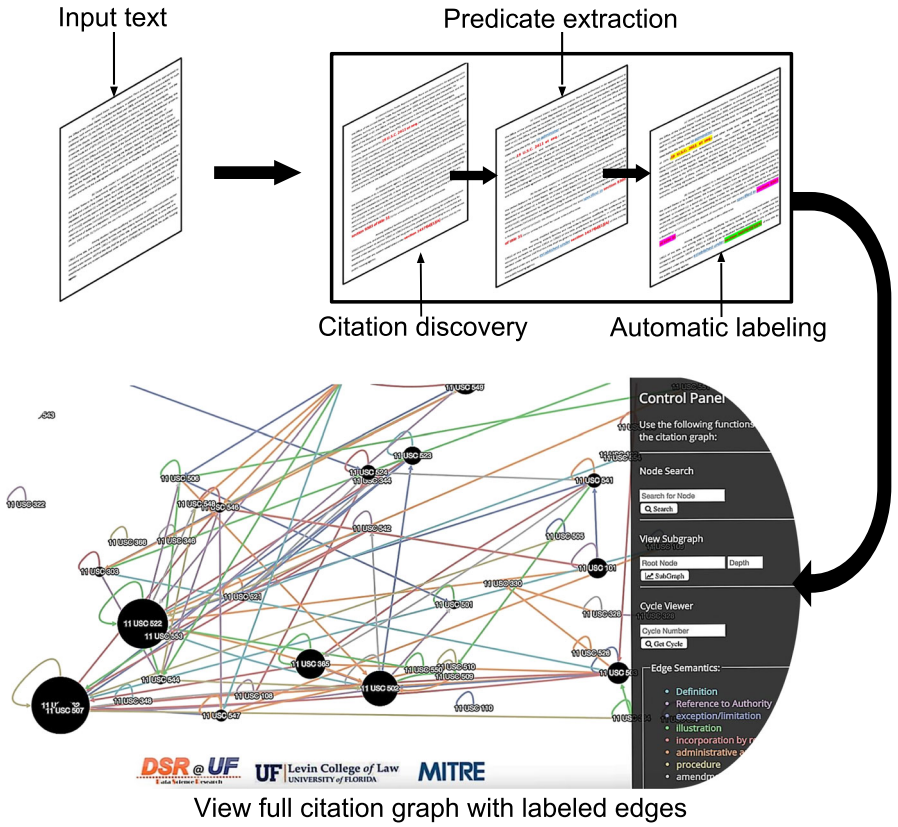

Fig. 1 Over view of the end-to-end system

illustrated using our graphical user interface where each edge type is colored according to its type. Figure 1 shows a diagram of the complete system.

5 Evaluation

In this section we evaluate the proposed gold standard label set used to capture the purpose of the citations, The annotated dataset, CRF model, and final Clustering Algorithm. We first briefly acquaint the reader with the dataset used, i.e. the US Code.

5.1 The dataset

The dataset used to demonstrate the use of our system is the US code, which is a consolidation of the general and permanent laws of the United States. There are in total over 130,000 different citations in the US Code. The collection of US Code used was taken from the online repository of Legal Information Institute of Cornell Law School.[1] There are over 29,000 distinct citations to statutes in the US Code, Code of Federal Regulations and other sources. These laws cite 26,417 distinct US codes with the US law code "42 USC 1395x" being cited the highest.

5.2 Designing the golden labels

We inspected many random provisions found in the US Code and proposed a primary set of labels that could capture the relations found there in. This labels along with a set of unlabeled citations from the US Code was then annotated by a group of expert human annotators (details of the annotation process are described in Sect. 5.3).

After analyzing the results of the beta round and integrating the feed back we got from the first round of annotations we updated the labels: Labels that were too close to be separated and caused confusion were merged; also, we expanded the labels by adding new labels found to be necessary by experts and annotators. We believe that the purpose of each citation can be effectively captured with this set of 9 labels.

- *Legal Basis* A relationship between a program/entity and the statute that is its legal basis.
- *Authority* A relationship under which one party is permitted to direct another party to perform an action.
- *Definition* A citation that directly defines the subject, brings a definition for a term used in the rule.
- *Example or illustrations* A citation to a rule that is used to introduce something chosen as a typical case or is defining the subject by illustrating/describing it.
- *Exception* A link between a rule and a set of circumstances where that rule doesn't apply.

[1] https://www.law.cornell.edu/uscode/text.

- *Criterion* A link from a conclusion to the "standard/criterion", but not how (not the procedure), of reaching that.
- *Limitation* A relationship between a description and a restriction on that.
- *Procedure* A link from an activity to a description of how that activity should be performed
- *Amended by/Amendment to* A relationship between two versions of the rule.

As we discuss in Sect. 5.3, the final round of annotations by the human experts confirmed the validity of this labels. The result is a label set long enough to cover almost all of the citations in the 1000 random samples from the US Code and also short enough for practical use.

We must note that designing a gold standard can be very use-case dependent and is an inherently difficult task. It requires field knowledge of the application and the related law. Our legal experts and our manual annotations imply that this set of labels are suitable for the US Code. But by no means do we indicate that this set of labels are globally applicable to all legal statutes and applications.

However, it is worth noting that although we don't claim global coverage, some other independent papers have reached a similar set of labels. As discussed in Sect. 2, Maxwell et al. (2012), Hamdaqa and Hamou-Lhadj (2011) suggest a very similar set of labels despite the fact that their focus is on the US Health Insurance Portability and Accountability Act (HIPAA) and the Gramm–Leach–Bliley Act (GLBA). In Breaux and Antón (2007), HIPAA Privacy Rule and the Telecommunications are used and the suggested set of labels are semantically similar to our proposed labels.

5.3 Annotation process

As we mentioned earlier, little prior work has been done on labeling citations. The few that did approach this problem (Maxwell et al. 2012; Hamdaqa and Hamou-Lhadj 2011, 2009; Breaux and Antón 2007) dont take an analytical approach or provide numerical results. Also, to the extent of our knowledge there is no annotated dataset available for the purposes of training and testing an automated system that clusters citations.

In this section, we explain how the dataset was generated. The final dataset is used to test the coverage of our gold set of labels and to apply machine learning paradigms for labeling the citations.

As mentioned in Alonso and Mizzaro (2012) the use of crowd-sourced options like amazon mechanical turk are usually a good option for similar tasks. But as our early analysis of the data made clear, this is a task that requires annotators to have a critical legal expertise and is a domain specific task. The manual annotators could experience problems like Logical Ambiguity, which would need legal expertise to be resolved (Rissland 1988). To decrease the misjudgments associated with non-expert annotators, a group of 7 paid Law graduate students with the guidance of a team of 2 legal experts were used to generate the dataset in two phases described below.

In order to design and complete the gold standard set of labels and to familiarize the annotators with the process, we first ran a small round of annotation and analyzed the results. In the beta phase, a total of 200 random citations were chosen from the US Code and were annotated. Each sample was annotated by 3 different analysts and majority voting was used to select the label for that citation. The annotators were able to either select one of the predefined labels in the label set provided or choose "New Label Necessary" if the particular edge didn't fall under any of the predefined set of labels.

After collecting the results from the beta round, we modified the pre-defined set of labels based on the feedback to compensate for: (1) The ambiguities in the definition of the labels and (2) The shortcoming in the coverage of the labels set.

In the final round of annotations, the same law students were given 1000 randomly selected paragraphs containing citations from the corpus of US code. Following the setup of the beta phase, each citation was annotated by 3 different people and majority voting was used to choose the label. We must note that to ensure a homogeneous annotation every data point was assigned to 3 randomly selected annotators but ensuring every annotator gets an approximately equal number of citations to annotate in total. The second round validated the gold standard as the manual annotators did not find a need to expand the labels set to accommodate for many citations, in fact only 1 out of 1000. This confirms that our label set covers the citations in the US Code very well.

Out of the 1000 annotated citations, two were discarded due to formatting issues of the text which was pointed out by the annotators. Table 1 presents the results obtained from the final round of annotations.

"Three different labels" counts the number of citations that received 3 different labels by the annotators. Some of the samples did not get a label from one or more of the annotators. "Unresolvable" counts the number of citations that could not be resolved with majority vote, i.e., one of the annotators didn't provide a label for it.

The high number of unresolved samples indicates that at least one of the annotators found it very hard to classify the citation and indicates the difficult nature

Table 1 Class label abbreviation and their count

Class name	Citation labels	Count
Le	Legal Basis	119
Cr	Criterion	81
De	Definition	53
Ex	Exception	38
Pr	Procedure	29
Am	Amendment	29
Au	Delegation of authority	22
Li	Limitation	22
Il	Example or illustration	1
–	New label necessary	1
–	Three different labels	111
–	Unresolvable	492

Citation types (classes) and the number of instances for each class obtained in the expert annotation process

of the problem even for a human expert annotator. Various factors contribute to the difficulty of this problem and thus many blank annotations. Such as the inherent difficulty of manually reading the law, multi-class nature of the citations, and also that the annotators were only looking at the citing text and did not have access to the actual text of the citation. For example:

> *...two or more districts may be merged as provided in section 2252(a)(2) of this title.*

Just by looking at the citing text it is not clear whether "section 2252(a)(2)" contains the legal basis of why 2 or more districts can be merged or the procedure of merging. This observations made it clear for us that it is very important to analyze the cited body too. Using the text of the cited provisions are out of the scope of our current work as we have not gathered annotated data, and we plan to address them in the future work.

From here on we will only focus on the citations that have been successfully classified by the annotators.

To measure the nominal agreement between annotators we use Cohen (1960). Kappa statistics is a score that measures the observed inter-annotator agreement offsetting the agreement by chance. Kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o is the probability of agreement by chance, i.e., expected proportion of agreement if labels were assigned by chance. p_e is the observed proportion of agreement of annotators. Usually values of 0.21–0.40 are considered as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial agreement (Landis and Koch 1977). Using (1) the inter-annotator coefficient calculated for our obtained dataset is 0.53.

The second round of annotations produced 394 labeled citations that served as the basis for the the Clustering algorithms explained in the following sections.

5.4 Predicate extraction

To find the predicate in the context of the citing provision, we used a linear-chain CRF. As we noted before, the predicate by definition will contain a lot of information about the citation and play an important role in finding the correct label.

Conditional random fields are probabilistic graphical models that very well capture the sequential property present in words in natural language (Lafferty et al. 2001). As stated before, the correlated sequential structure of the words in a predicate can be well captured with this type of graphical models.

We replace the whole span of the target citation text, for which we intend to find the predicate, with a unique character sequence not present in the rest of the corpus, i.e, `C1CITE`. This will make it easier for the CRF to recognize the the citation and work with it as a single phrase. To mark the span of each predicate we used the standard Begin/In/Out (BIO) encoding for tagging the predicate chunks and the other tokens. To evaluate the performance of this model, we applied the system to a

dataset of 1000 citations and their corresponding predicates.[2] To train the CRF we used the features describe in 4.2.

We performed a tenfold cross validation and presented the performance results in Table 2.

5.5 Unsupervised clustering accuracy

As mentioned before Sadeghian et al. (2016), after extracting each citation's predicate we used word2vec (Mikolov et al. 2013a, b) to represent each word in the predicate as a vector in a 300-dimensional space. To further simplify the clustering, we correspond each predicate with the average of the vectors representing each of the words in that predicate. Although this averaging results in a loss of information, but due to the properties of the embedding method used most of the meaning in the predicate is still preserved.

To cluster the data we used k-means classification and clustered the whole US Code using 15 cluster centers. Note that there is a relatively large number of labels and there is no guarantee that each form exactly one cluster in the projected space. For this reasons, we use more cluster centers to capture the spread as much as possible. This might slightly over-fit or even decrease the accuracy, but its effects are negligible compared to the relatively large dataset and number of labels.

To evaluate the performance of our clustering algorithm, we use the annotated dataset obtained from the human expert annotators. Each cluster is labeled according to the label of the closest point to the center of the cluster. We present the classification accuracy and the confusion matrix in Table 3.

As shown in the results Table 3, clustering the data with a vanilla K-means using 'average predicate word-vectors' predicts relatively well for some labels (Amendment, Definition, Legal Basis) but quite poorly on some (Procedure). This can be due to the lack of information in the predicates, ambiguous labels or due to inherent inadequacy of K-means. To measure the effect of the learning algorithm, the next section studies the use of two supervised learning methods on the same features and labels.

5.6 Supervised clustering accuracy

As we explained in Sect. 5.3, identifying the correct label for citations is a complicated task even for expert human annotators. This is partly due to the fact that the purpose of each citation is not always clearly expressed with a single label. Also, because there is a gray area between each class that the correct label for citations are uncertain and could potentially be any of the labels unless the cited provision is also examined. We set up the problem of classifying the citations as a multiclass classification problem. Due to certain commonalities shared between these classes, we evaluate the highest 2 predicted scores (top-2 method) to check if the classifier

[2] This dataset was also obtained during the annotation process, but lacked a semantic label for the citations. This reduces the chances of over fitting because the predicate extraction is learned on a different dataset than the dataset used for training the label classifier.

Table 2 Predicate extraction performance

	Prec.	*Recall*	*F1*	*Support*
B_PRD*	0.91	0.84	0.875	100
I_PRD	0.93	0.88	0.898	119
O	0.99	0.99	0.998	6518

*Following BIO encoding the beginning of a predicate is tagged with B_PRD, any other word in the predicate span is tagged with I_PRD and any word that is not a part of the predicate is tagged with O

Table 3 Empirical confusion matrix

	Am.	Au.	Cr.	De.	Il.	Ex.	Le.	Li.	Pr.
Am.	25	0	1	0	0	2	1	0	0
Au.	0	3	6	0	0	0	13	0	0
Cr.	0	0	47	0	0	4	29	1	0
De.	0	0	4	44	0	1	3	0	0
Il.	0	0	0	0	0	0	0	1	0
Ex.	0	0	6	0	0	22	6	4	0
Le.	2	0	9	2	0	7	98	1	0
Li.	0	0	8	0	0	2	3	9	0
Pr.	0	0	10	0	0	6	13	0	0

calls for the right classification of the citation among its top-2 most confident predictions.

We evaluate our method on the collected dataset. Following the setup of our previous experiment: each predicate is mapped to its average 300-dimensional word embedding. To minimize any over-fitting the dataset is split into a 0.33/0.77 test/train set. We use one-against-one approach and all hyper-parameters and results are obtained from a fivefold cross validation.

Top-2 accuracy is computed using the distance of the samples to the separating hyperplanes as a scoring function. We report prediction accuracies for SVM and Logistic Regression in Table 4.

It was apparent that the top-2 classifiers would yield a higher accuracy. However, the increase in accuracy by chance would have been around 4.5%. The huge jump in accuracies shows that the first 2 guesses are correct much more often, 79.8% accuracy versus 64.3%, This further validates our intuitions. Table 5 shows a few examples of the true label and the corresponding top two guesses for 15 random citations.

Table 4 SVM and Logistic Regression classification accuracies for top-1 and top-2 for

	Top-1	Top-2
Logistic Regr. (l_2 penalty)	0.61	0.75
Linear-SVM (c = 1)	0.64	0.79

Table 5 Example of true labels and the corresponding top-2 labels for SVM

Class label	Top-1	Top-2
LegalBasis	LegalBasis	Criterion
Criterion	Criterion	Definition
Authority	LegalBasis	Criterion
Authority	Authority	Exception
Limitation	LegalBasis	Limitation
Definition	Definition	LegalBasis
Criterion	LegalBasis	Criterion
Criterion	LegalBasis	Criterion
LegalBasis	LegalBasis	Criterion
Definition	Definition	LegalBasis
Criterion	Criterion	Exception
LegalBasis	LegalBasis	Authority
LegalBasis	LegalBasis	Criterion
LegalBasis	LegalBasis	Criterion
LegalBasis	Limitation	LegalBasis

6 Conclusion and future work

We presented an automated system that determines the purpose behind a citation. This enables lawyers and policymakers to better analyze the relation between different laws or users to find the necessary regulations much easier.

Our system has three main parts. We first automatically extract the citations from the document, then find an informative expression from the text related to that citation which we call the predicate. Using Natural Language Processing (NLP) and Machine Learning (ML) techniques we then label the citation into one of the predefined set of citation types.

Our contributions in this paper are three-fold. We propose a gold standard label set that almost all the citations in the US Code can be categorized according to it and verified its coverage in a manual experiment by a group of experts. We also produced a dataset of 394 annotated citations from the US code that can be used for future research on this topic. Finally, we built a fully automated system for semantic labeling of the edges over a legal citation graph, which we plan to open source.

In future work, we will have a more in-depth analysis of the results from annotation process and the accuracy of a human expert. We further plan to use advanced machine learning techniques to increase the accuracy of our system by

using the whole context related to the citing and also use the text in the cited provisions.

Acknowledgements We thank two anonymous reviewers for their insightful feed back, which helped us improve this manuscript. In addition the authors would like to thank Vironica I Brown, Roman Diveev, Max Goldstein, Eva L Lauer, Nicholas W Long, Paul J Punzone and Joseph M Ragukonis for their contributions in the annotation process. We would also like to thank Benjamin Grider for his help in designing the graphical user interface for our system.

References

Adedjouma M, Sabetzadeh M, Briand LC (2014) Automated detection and resolution of legal cross references: approach and a study of luxembourg's legislation. In: Requirements Engineering Conference (RE), 2014 IEEE 22nd International. IEEE, pp 63–72

Alonso O, Mizzaro S (2012) Using crowdsourcing for TREC relevance assessment. Inf Process Manag 48:1053–1066

Amir Sadeghian A, Alahi A, Savarese S (2017) Tracking the untrackable: learning to track multiple cues with long-term dependencies. arXiv preprint arXiv:1701.01909

Ashley K, Bjerke E, Potter M, Guclu H (2014) Statutory network analysis plus information retrieval. In: Proceedings of Second Workshop on Network Analysis in Law at the 27th Annual Conference on Legal Knowledge and Information Systems. NAil, pp 1–7

Association HLR (1996) The bluebook: a uniform system of citation. Harvard Law Review Association, Cambridge

Bird S, Klein E, Loper E (2009) Natural language processing with Python. O'Reilly Media Inc, Sebastopol

Branting LK (2017) Data-centric and logic-based models for automated legal problem solving. Artif Intell Law 25(1):5–27

Breaux TD, Antón AI (2007) A systematic method for acquiring regulatory requirements: a frame-based approach. In: RHAS-6, Delhi, India

Cao Z, Yu S, Ouyang B, Dalgleish F, Vuorenkoski A, Alsenas G, Principe J (2017) Marine animal classification with correntropy loss based multi-view learning. arXiv preprint arXiv:1705.01217

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

Cornell Law School US Code. https://www.law.cornell.edu/uscode/text

de Maat E, Winkels R, van Engers T (2006) Automated detection of reference structures in law. In: van Engers TM (ed) Legal knowledge and information systems. Jurix 2006: the nineteenth annual conference. Frontiers in artificial intelligence and applications, vol 152. IOS Press, pp 41–50

de Maat E, Winkels R, van Engers T (2009) Making sense of legal texts. Form. Linguist. Law 212:225

Galgani F, Hoffmann A (2010) Lexa: towards automatic legal citation classification. In: AI 2010—Advances in Artificial Intelligence. Springer, Berlin, pp 445–454

Glaser B, Strauss A (1967) The discovery grounded theory: strategies for qualitative inquiry. Aldin, Chicago

Hamdaqa M, Hamou-Lhadj A (2009) Citation analysis: an approach for facilitating the understanding and the analysis of regulatory compliance documents. In: Sixth International Conference on Information Technology—New Generations, 2009. ITNG'09. IEEE, pp 278–283

Hamdaqa M, Hamou-Lhadj A (2011) An approach based on citation analysis to support effective handling of regulatory compliance. Future Gener Comput Syst 27:395–410

Harrington WG (1984) Brief history of computer-assisted legal research. Law Libr J 77:543

Jain A, Lopez-Aguilera E, Demirkol I (2017) Mobility management as a service for 5G networks. arXiv preprint arXiv:1705.09101

Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 282–289

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Maxwell JC, Antón AI, Swire P, Riaz M, McCraw CM (2012) A legal cross-references taxonomy for reasoning about compliance requirements. Requir Eng 17:99–115

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems 26. Curran Associates, Inc., pp 3111–3119

Mollalo A, Alimohammadi A, Shirzadi M, Malek M (2015) Geographic information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, north-east of Iran. Zoonoses Public Health 62:18–28

Neale T (2013) Citation analysis of canadian case law. J. Open Access L. 1:1

Pollman T, Kane LA (2000) ALWD citation manual: a professional system of citation. UNLV School of Law, Las Vegas

Prakken H (1993) A logical framework for modelling legal argument. In: Proceedings of the 4th International Conference on Artificial Intelligence and Law. ACM, pp 1–9

Rissland E (1988) Artificial intelligence and legal reasoning: a discussion of the field and gardner's book. AI Mag 9:45

Rodrıguez M, Goldberg S, Wang DZ (2016) Consensus maximization fusion of probabilistic information extractors. In: Proceedings of NAACL-HLT, pp 1208–1216

Roitblat HL, Kershaw A, Oot P (2010) Document categorization in legal electronic discovery: computer classification versus manual review. J Am Soc Inf Sci Technol 61:70–80

Sadeghian A, Lim D, Karlsson J, Li J (2015) Automatic target recognition using discrimination based on optimal transport. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 2604–2608

Sadeghian A, Sundaram L, Wang D, Hamilton W, Branting K, Pfeifer C (2016) Semantic edge labeling over legal citation graphs. In: LTDCA

Sharghi A, Laurel JS, Gong B (2017) Query-focused video summarization: dataset, evaluation, and a memory network based approach. arXiv preprint arXiv:1707.04960

Sutton C, McCallum A (2006) An introduction to conditional random fields for relational learning, vol 2. Introduction to statistical relational learning. MIT Press

Tran OT, Ngo BX, Le Nguyen M, Shimazu A (2014) Automated reference resolution in legal texts. Artif Intell Law 22:29–60

Winkels R, Boer A, Vredebregt B, van Someren A (2014) Towards a legal recommender system. In: JURIX

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