CS447 Literature Review: Automating Regulatory-Related Document Search in Regulated Businesses

David Silverman,
davids17@illinois.edu

November 22, 2023

Abstract

Companies are required to comply with government regulations. To evidence this compliance, companies search across multiple datasets for documents related to regulations. This information retrieval task could be made less manual through the use of NLP techniques. This paper reviews research papers with potential solutions to this search problem and considers if they might prove appropriate for this use case.

1 Introduction

Businesses are required to comply with some amount government regulations. For example in industries such as energy, pharmaceuticals, health care, communications, social media, and especially so in finance.

For example, at a large financial firm such as JP Morgan Chase, it must comply with literally thousands of pages of rules from regulators such as the US Federal Reserve, US Consumer Financial Protection Bureau, Securities and Exchange Commission, US Office of the Comptroller of the Currency, Financial Industry Regulatory Authority, and many more including US States and regulators around the world.

To evidence compliance with these rules, corporations are reviewed by multiple groups: externally by government regulators and internally by Compliance Testing and Audit departments. In all cases, the process is the same: the examiners request documentation that shows the company has policies and procedures in place for a given regulation and that these are designed appropriately and functioning effectively.

This documentation search, or alternately information retrieval task, is non-trivial because the required information is scattered across multiple datasets and systems. An average question at a large financial firm will take 2 to 3 months of time, directly involve 5 to 10 staff, and impact hundreds of other employees who have relevant information. Figure 1 provides a view of the process.

The impact of failing to answer an inquiry can be financial penalties, removal of corporate officers, or loss of a banking license. In terms of raw monetary cost for non-compliance: there have been over $300 billion in fines and settlements with banks since the year 2000, just in the United States. (Source: https://violationtracker.goodjobsfirst.org/top-industries

Therefore finding ways to use NLP to automate the search for documents related to regulations is well worth the effort. This paper will examine various methods for connecting company documents to their related regulations and discuss how these methods could be potentially useful for solving this problem and where there are challenges.

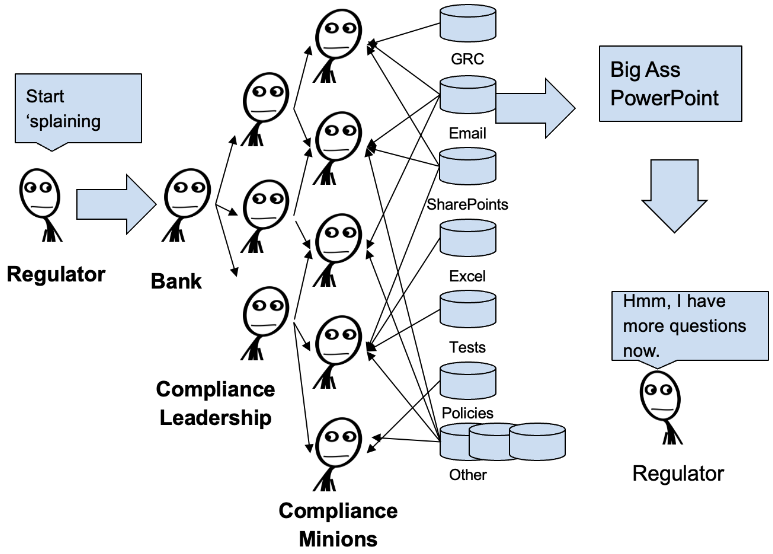


Figure 1: The Bank Regulatory Exam Process
From my book: *Stop Harming Customers: A Compliance Manifesto*
https://www.businessexpertpress.com/books/stop-harming-customers-a-compliance-manifesto/

2 Background: Regulatory-Related Document Retrieval Use Case

Regulation is hierarchical with statute laws being composed of regulatory codes (Penal/Criminal, Civil, State/Local). Figure 2 illustrates this structure. Ultimately, we want to retrieve and link to information at the smallest appropriate lexical scope to maximize relevance and applicability.

As an example, here's a specific regulatory rule to show what we need to retrieve and how NLP might be valuable. US Federal banking law requires that if a customer performs a transaction of more than $5,000, and the bank suspects the money laundering, the bank must file a *suspicious activity report* (SAR) with the government within 30 days.

The specific citation is: 12 CFR 21.11, where CFR is the US Code of Federal Regulations. Figure 3 contains a snippet of this regulation edited to focus on the most relevant sections:

To evidence compliance with this rule, the bank might want to find any of the following documents related to SARs:

- policies
- procedures
- records of transactions
- training
- checklists
- metrics
- software specs
- projects reports
- meeting minutes
- emails
- spreadsheets
- tests or examinations
- open issues

A simple keyword search of the company intranet for 12 CFR 21.11, SAR, anti-money laundering, $5,000 or other terms highlighted in Figure 3 will yield many false positives and miss false

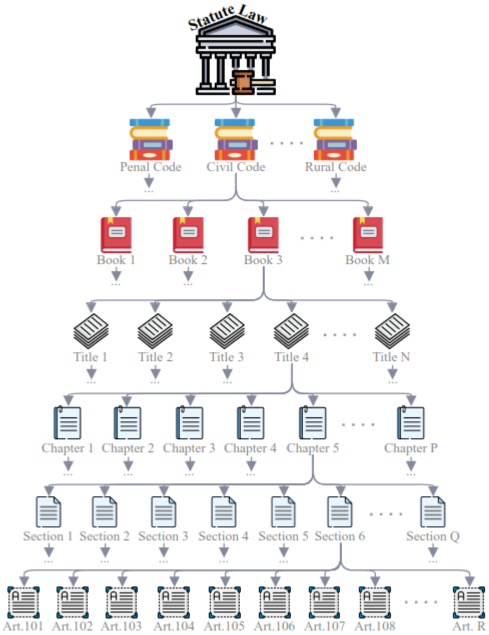


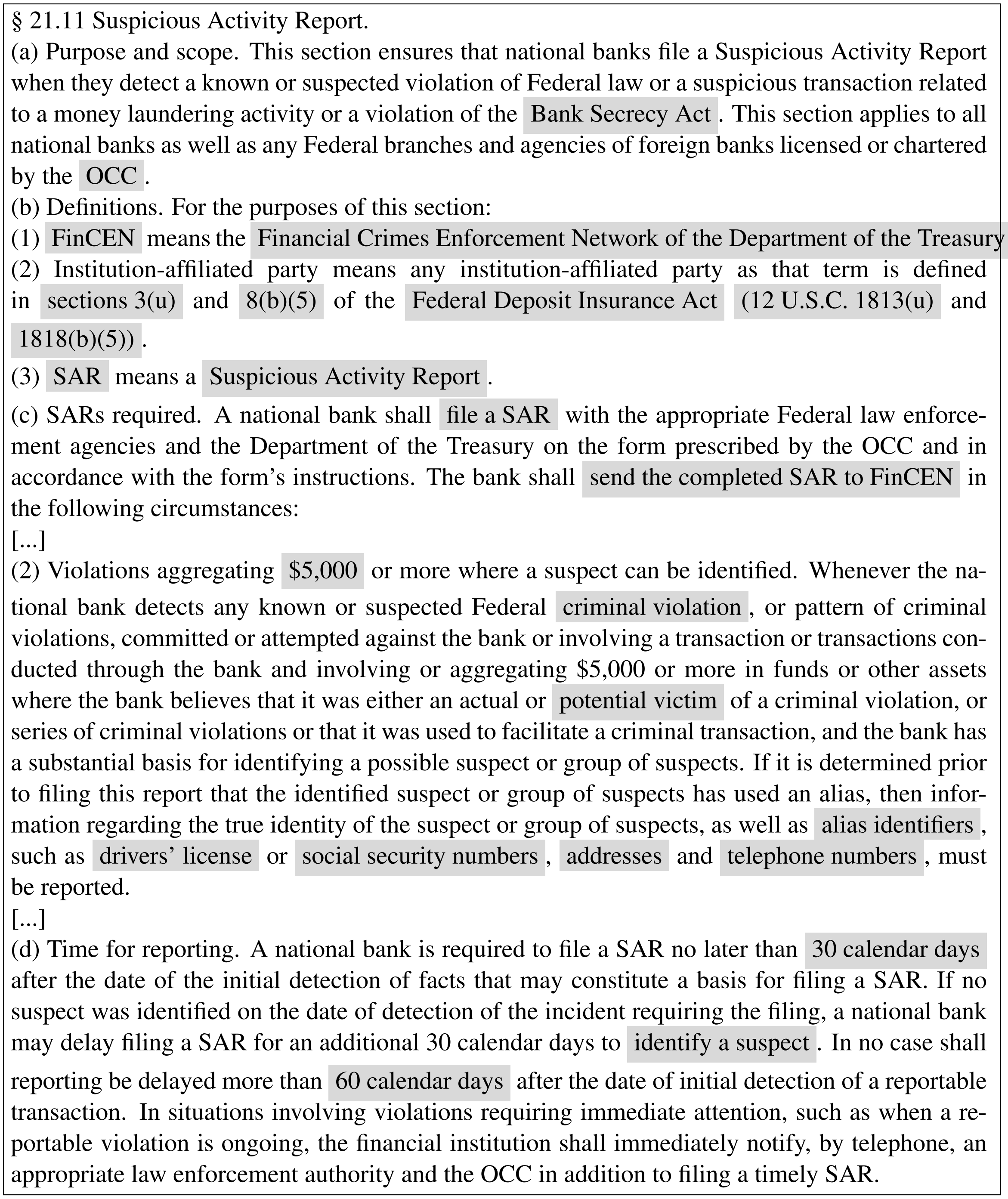
Figure 2: Hierarchical structure of regulation (Louis et al., 2023)

negatives. This is because company documents typically do not directly reference the regulation—as the context of the regulation is assumed by the authors much like fish assuming they are in water. Additionally, terms can have multiple meanings: $5,000 could be a SAR trigger or the balance needed to qualify for a new credit card; the requirement for social security numbers is also part of tax filing requirements; and "30 calendar days" will occur in many unrelated documents.

3 List of Papers

Following are the papers that will be reviewed and discussed. For each, I provide a summary of the paper, including a description of the methods used, and then a two-part analysis. First, why the paper's approach may work for regulatory-related documents, and second, why the paper's approach may *not* work for regulatory-related documents. This is all supported by images and results from the source papers.

- Paper 1: Cross-Document Co-Reference Resolution using Sample-Based Clustering with Knowledge Enrichment (Dutta and Weikum, 2015)
- Paper 2: CDˆ2CR: Co-reference resolution across documents and domains (Ravenscroft et al., 2021)
- Paper 3: Legal Linking: Citation Resolution and Suggestion in Constitutional Law (Shaffer and Mayhew, 2019)
- Paper 4: Finding the Law: Enhancing Statutory Article Retrieval via Graph Neural Networks (Louis et al., 2023)
- Paper 5: Prototype-Based Interpretability for Legal Citation Prediction (Luo et al., 2023)



§ 21.11 Suspicious Activity Report.

(a) Purpose and scope. This section ensures that national banks file a Suspicious Activity Report when they detect a known or suspected violation of Federal law or a suspicious transaction related to a money laundering activity or a violation of the Bank Secrecy Act . This section applies to all national banks as well as any Federal branches and agencies of foreign banks licensed or chartered by the OCC .

(b) Definitions. For the purposes of this section:

(1) FinCEN means the Financial Crimes Enforcement Network of the Department of the Treasury .

(2) Institution-affiliated party means any institution-affiliated party as that term is defined in sections 3(u) and 8(b)(5) of the Federal Deposit Insurance Act (12 U.S.C. 1813(u) and 1818(b)(5)) .

(3) SAR means a Suspicious Activity Report .

(c) SARs required. A national bank shall file a SAR with the appropriate Federal law enforcement agencies and the Department of the Treasury on the form prescribed by the OCC and in accordance with the form's instructions. The bank shall send the completed SAR to FinCEN in the following circumstances:

[...]

(2) Violations aggregating $5,000 or more where a suspect can be identified. Whenever the national bank detects any known or suspected Federal criminal violation , or pattern of criminal violations, committed or attempted against the bank or involving a transaction or transactions conducted through the bank and involving or aggregating $5,000 or more in funds or other assets where the bank believes that it was either an actual or potential victim of a criminal violation, or series of criminal violations or that it was used to facilitate a criminal transaction, and the bank has a substantial basis for identifying a possible suspect or group of suspects. If it is determined prior to filing this report that the identified suspect or group of suspects has used an alias, then information regarding the true identity of the suspect or group of suspects, as well as alias identifiers , such as drivers' license or social security numbers , addresses and telephone numbers , must be reported.

[...]

(d) Time for reporting. A national bank is required to file a SAR no later than 30 calendar days after the date of the initial detection of facts that may constitute a basis for filing a SAR. If no suspect was identified on the date of detection of the incident requiring the filing, a national bank may delay filing a SAR for an additional 30 calendar days to identify a suspect . In no case shall reporting be delayed more than 60 calendar days after the date of initial detection of a reportable transaction. In situations involving violations requiring immediate attention, such as when a reportable violation is ongoing, the financial institution shall immediately notify, by telephone, an appropriate law enforcement authority and the OCC in addition to filing a timely SAR.

Figure 3: 12 CFR 21.11 (from `https://www.ecfr.gov/current/title-12/section-21.11`) with some relevant entities highlighted

4 Paper 1: Cross-Document Co-Reference Resolution using Sample-Based Clustering with Knowledge Enrichment (Dutta and Weikum, 2015)

The paper Dutta and Weikum (2015) presents the author's "CROCS framework" for performing unsupervised cross-document co-reference resolution. That is, a method for finding references to the same entity across multiple documents. CROCS use what the authors call "semantic summaries" (*sensums*) which are made by combine entities extracted from documents with semantic relationships from knowledge bases (*KBs*). The authors use Freebase (which has been replaced by Wikidata `https://www.wikidata.org/` after the paper was written) and Yago (`https://yago-knowledge.org/`) as the KBs and Stanford's CoreNLP NER tagger to extract named entities from the documents.

The paper imagines the text "Hillary lived in the White House and backed Bill despite his affairs." within a sample to document. To find related documents, we need to know that Hillary refers to "Hillary Clinton", "President Clinton's wife", "Senator Clinton" and so on, and to disambiguate "Sir Edmund Hillary" so that we return documents about American politics and not British mountaineering.

The paper does this by connecting the KB content for nearby entities e.g., "White House", with the KBs for Hillary and selecting the highest ranked relationships based on the number of 2-hop distance co-occurrence of the terms with in the KB. This is referred to as *knowledge enrichment* per the paper title.

CROCs then performs this enrichment across all *local mention groups* in each document. A local mention group is a set of named entities (extracted by Stanford CoreNLP) within a limited distance in the document. Note: I could not find anywhere in the paper how local mention groups are determined. However, there is an illustration in the paper that makes the idea clearer, see Figure 4.

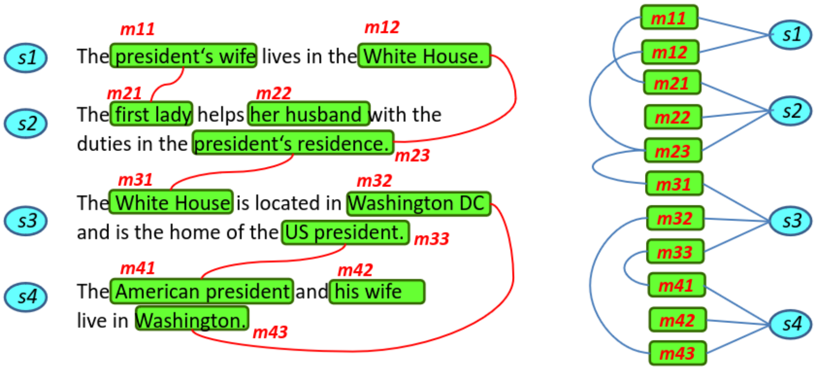

s1
m11
m12
The president's wife lives in the White House.
s2
m21
m22
The first lady helps her husband with the
duties in the president's residence.
m23
s3
m31
m32
The White House is located in Washington DC
and is the home of the US president.
m33
s4
m41
m42
The American president and his wife
live in Washington.
m43
m11
m12
m21
m22
m23
m31
m32
m33
m41
m42
m43
s1
s2
s3
s4

Figure 4: Example of local mention group (Dutta and Weikum, 2015)

This produces a feature vector that allows for two entities to be scored for similarity. The equation is:

$$sim(G_1, G_2) = \alpha \times sim_{BoW}(G_1, G_2) + (1-\alpha) \times sim_{KP}(G_1, G_2)$$

Where G_1 and G_1 are mention groups of entities, sim_{BoW} is the similarity of the two groups computed by TF/IDF looking only at nearby words and sim_{KP} is the similarity (also by TF/IDF) looking at the extracted named entities and related KB entities. α is a tunable hyper-parameter to weight one or the other similarity score.

Given this score, a clustering algorithm (spectral clustering) is then used to partition the entire mention group into small, related sets of entities that comprise the final cross-document entity links.

4.1 Analysis: Why this may work for regulatory-related documents

Much of the challenge of corporate and regulatory documents is frequent reuse of the same term with different meaning. The ability to resolve entities based on their semantic meaning could aid retrieving documents, such as policies and procedures, related to our example for suspicious activity reports in section section 2, while avoiding documents that talk about other kinds of reports, would be very useful.

The paper's key idea of connecting terms to KB to provide semantic context, and thereby overcome term ambiguity, is very interesting. In both regulations and corporate documents, there are often definitions of terms that could serve as the basis of a company/legal knowledge graph/ontology. And there are a variety of legal ontologies available as a starting point. (`https://github.com/Liquid-Legal-Institute/Legal-Ontologies`)

4.2 Analysis: Why this may not work for regulatory-related documents

There are several weaknesses in the paper. First there is a reliance on named entities extractable by Stanford's NER algorithm. However, in the regulatory use case, the entities to disambiguate are nouns such as "customer" where the meaning might be a person or a corporation. This would not be identified using the method in the paper.

Secondly, while the method worked well on small datasets of less than 5,000 web pages, the authors tests on a large dataset of 1.8 million news articles from the New York Times achieved a F1 score of only 62.2%, see Figure 5. This indicates the approach would likely need to be greatly adapted to be effective in the regulatory document use case.

Dataset	**Freebase**		
	P(%)	*R(%)*	*F1*
WePS-2	86.3	82.1	83.9
NYT	59.8	64.9	62.2

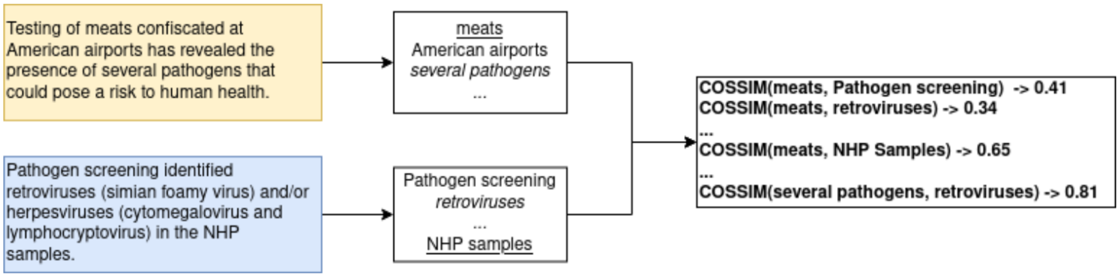
Figure 5: Paper 1 precision and recall (Ravenscroft et al., 2021)

5 Paper 2: CD^2CR: Co-reference resolution across documents and domains (Ravenscroft et al., 2021)

My second paper, (Ravenscroft et al., 2021), also aims to resolve entity co-references across multiple documents. The similarity with the first paper is no accident; the authors reference (Dutta and Weikum, 2015) from 6 years earlier as a primary sources. However, this paper differs in two primary aspects.

First it aims to connect documents across different domains. Whereas (Dutta and Weikum, 2015) looked for entity co-references in a single domain, e.g., NY Times news articles, this paper looks for connections between news articles and scientific papers. This is harder because the content of words and their context neighbors differ across the domains. For example the scientific statement "convalescent plasma derived from donors" can become "blood from recovered patients" in the news. Further examples are in Figure 6.

Secondly, rather than using knowledge bases and clustering, it makes use of modern neural transformer methods (i.e., BERT) that have been invented since the earlier research was done. To



Testing of meats confiscated at American airports has revealed the presence of several pathogens that could pose a risk to human health.

Pathogen screening identified retroviruses (simian foamy virus) and/or herpesviruses (cytomegalovirus and lymphocryptovirus) in the NHP samples.

meats
American airports
several pathogens
...
Pathogen screening
retroviruses
...
NHP samples
COSSIM(meats, Pathogen screening) -> 0.41
COSSIM(meats, retroviruses) -> 0.34
...
COSSIM(meats, NHP Samples) -> 0.65
...
COSSIM(several pathogens, retroviruses) -> 0.81

Figure 6: Co-reference between news summary and scientific abstract. Pairs are shown with same formatting (underline, italic) (Ravenscroft et al., 2021)

provide the *gold standard* for BERT training, the researchers created software for human annotators to "identify co-referent mentions between the scientific paper abstract and a summary of the news article."

These annotations are then used to fine tune a feed-forward model. The authors evaluate both a RoBERTa encoder, a version of BERT trained on a large corpus of text, and SciBERT, which is a version of BERT pretrained on scientific literature.

5.1 Analysis: Why this may work for regulatory-related documents

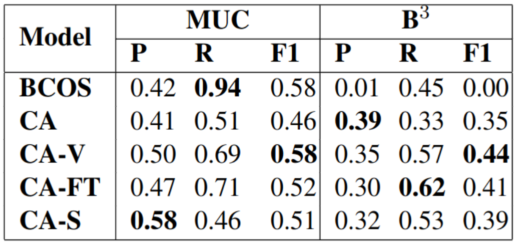
Connecting across domains is especially relevant to this use case. The law and corporate documents, much like news articles and scientific papers, are written in entirely different styles. Additionally, the open source annotation pipeline developed by the authors provides a straightforward way to gather training data that is adaptable to other use cases. (`https://github.com/ravenscroftj/cdcrtool`)

5.2 Analysis: Why this may not work for regulatory-related documents

Sadly, the results of the research were not promising. The authors tested 5 different models ranging from baseline with unmodified BERT and an unsupervised RoBERTa to the models tuned with the annotation data. The highest resulting recall, precision and F1 scores were spread across all of them, with no clear winner. See Figure 7. In fact, baseline BERT managed the best recall (0.94 MUC) and none of the F1 measures were above 0.58. Given that for the regulatory-related use case, recall is more important than precision—because we want to be sure not to miss any laws or corporate documents—this implies that baseline BERT rather than any of the paper's models would be best suited.

The authors suggest that this is two-fold. One, annotation is difficult and requires a lot of domain expertise. The "three university-educated human annotators" struggled with relating jargon from highly technical scientific abstracts to general news articles. I would agree this is also the case for regulations and corporate documents. Therefore creating the annotations is going to be just as difficult, and require as much expertise, as creating a knowledge base as per (Dutta and Weikum, 2015).

Two, the models do not incorporate any lexical knowledge and there is "a significant overlap between co-referent and non-co-referent cosine similarities" meaning that the classification layer



Model	**MUC**			**B^3**		
	P	**R**	**F1**	**P**	**R**	**F1**
BCOS	0.42	**0.94**	0.58	0.01	0.45	0.00
CA	0.41	0.51	0.46	**0.39**	0.33	0.35
CA-V	0.50	0.69	**0.58**	0.35	0.57	**0.44**
CA-FT	0.47	0.71	0.52	0.30	**0.62**	0.41
CA-S	**0.58**	0.46	0.51	0.32	0.53	0.39

Figure 7: Paper 2 precision and recall
MUC and B^3 Accuracy. BCOS = baseline BERT, CA = baseline RoBERTa, CA-V = RoBERTA + CD^2CR annotations no-pretrain, CA-FT = use CD^2CR annotations with pretrain, CA-S = SciBert no pretrain) (Ravenscroft et al., 2021)

does not end up linking entities across domains. This would, of course, also be a problem for regulatory document search.

6 Paper 3: Legal Linking: Citation Resolution and Suggestion in Constitutional Law (Shaffer and Mayhew, 2019)

My third paper is Shaffer and Mayhew (2019) which focuses on linking paragraphs to legal text. Specifically, from court cases to the US Constitution. To do this, the authors built a training set of ∼328k unique paragraphs from the Cornell Legal Information Institute and scanned for hyperlinks to Supreme Court cases resulting in ∼11k links.

Next they created ∼100 rule strings such as "Seventh Amendment", "7th Amendment", "free speech", "due process" and so on, each linked to the relevant Constitutional section. Finally, to "encourage the model to move beyond these trival patterns" they created a "stripped" dataset where they randomly removed links and rule strings from half the training set.

They then applied and compared three different approaches:

1. **Rule-based** Rules captured by annotation of the training data are used as the labeling strategy for the entire data set.

2. **Linear model** Logistic regression using unigrams, bigrams, and trigrams with a linear multi-label classifier that is a variant of a *binary relevance framework*, which ignores relationships between labels and builds a separate classifier for each. Notationally, for each label C there is $h(d) \to \{h_i(d)\}_{i=1}^{|C|}$ where $h_i(d) \to 0, 1$.

3. **Neural mode** Using *doc2vec* the model combines BERT embeddings of the source paragraph and the related Constitution sections as a document. See Figure 8.

To evaluate the models, they hand annotated constitutional references from 1,241 paragraphs from five Supreme Court cases to provide a gold standard.

6.1 Analysis: Why this may work for regulatory-related documents

This paper is directly applicable to finding regulatory-related documents as it works with legal texts: case law and Supreme Court decisions. Unfortunately, this is about the extent of the good news.

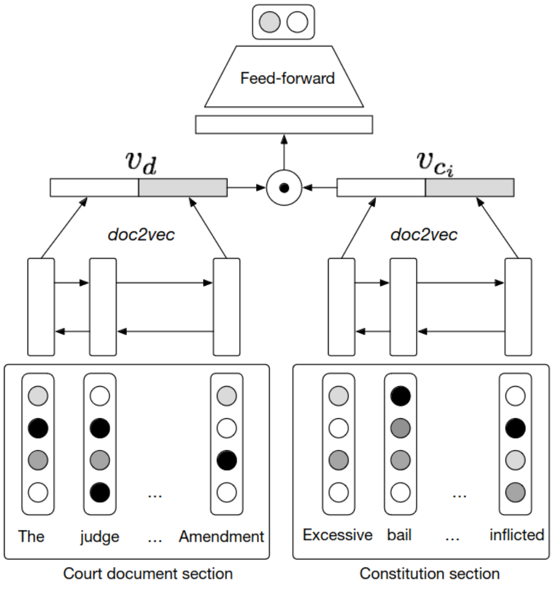


Figure 8: Neural Model (the ⊙ is element-wise multiplication) (Shaffer and Mayhew, 2019)

6.2 Analysis: Why this may not work for regulatory-related documents

The Constitution is very small, $< 10^3$ words, (`https://wordcounter.net/`) in comparison to the corpus of regulations, $> 10^6$ words, (`https://arxiv.org/pdf/1003.4146.pdf`) and is therefore likely a poor basis for developing a model that can link arbitrary sections of regulations to corporate documents.

Moreover, the results are not encouraging. While the stripped nerual model achieved an F1 of 64.8, this is not much better than the simple rules-based approach at 62.2, yet with much worse precision. See Figure 9.

The authors theorize that this is partially due to the "long tail" of infrequent terms, per Zipf's law, and that there is not enough training data for these rare terms.

Model	P	R	F1
Rule-based	**91.8**	47.0	62.2
Linear (original)	79.0	45.8	58.0
Linear (stripped)	68.3	54.3	60.5
Neural Network (original)	82.1	46.8	59.6
Neural Network (stripped)	76.5	**56.2**	**64.8**

Figure 9: Paper 3 precision and recall (Shaffer and Mayhew, 2019)

7 Paper 4: Finding the Law: Enhancing Statutory Article Retrieval via Graph Neural Networks (Louis et al., 2023)

My fourth paper is Louis et al. (2023), in which the authors describe *statutory article retrieval* (SAR)[1] as the task of answering legal questions such as “Who should pay for the construction of the common wall?” with relevant regulatory citations. They then propose a *graph-augmented dense statue retriever* (G-DSR).

This model, which is depicted in Figure 10, has two primary components:

1. *Dense statute retriever* (DSR)
2. *Legislative graph encoder* (LGE)

The DSR uses an encoder to create embeddings for a given query and a dataset of statutory articles. The query embedding is a straightforward application of a pretrained BERT model. The statues, however, are longer than the $n_{max} = 512$ token limit of BERT. To overcome this, the authors use a *hierarchical article encoder* to split the articles into passages with $n \leq 512$ which are passed through BERT and then through a transformer layer with learnable passage position embeddings and then pooled into a single embedding.

The query and statutory articles are then are scored for relevance via a similarity function:

$$s(q,a) = sim(E_Q^\theta(q), E_Q^\phi(a))$$

where $E_Q^\theta(q), E_Q^\phi(a) \in \mathbb{R}^d$ are the query and article embedding respectively and *sim* is either cosine or dot-product.

The DSR is trained on a combination of positive and negative examples. The negative examples are extracted from by selecting articles paired with non-associated queries in the training set and by also generating likely negatives via BM25, which is a term-based retrieval method that is a variant of TF/IDF.

The LGE “aims to enrich article representations...by fusing information from a legislative graph.” This is accomplished by building a graph of nodes, one for each article, with edges linking them together into sections by a directed acyclic graph, in this case a tree, as depicted in Figure 2. The nodes are initialized with the article embeddings from the DSR and then updated graph neural network (GNN) that uses dynamic attention to weight the strength of the connection to neighboring nodes. To reduce computational complexity, the authors only look at neighbors L-hops distant. L is set to the number of layers of the GNN, which they set to 3 after testing $L \in [1,2,3,4]$.

7.1 Analysis: Why this may work for regulatory-related documents

This paper’s topic and approach is directly related to locating regulations related to corporate documents. If we replace the query about “common walls” with a the text of a corporate procedure, and apply the same transformer-based approach for processing this procedure into $n \leq 512$ tokens, we should be able to directly test their methods for regulatory-related document search. Further, corporate documents, like statutory articles, have a hierarchical structure that could work with the same GNN approach used in the paper’s LGE, providing another potential avenue to explore.

Of course, this is all dependent on how effective the G-DSR approach was. To evaluate recall and accuracy, the authors used the BSARD dataset of “1,100+ French native legal questions

[1] Not to be confused with Suspicious Activity Report from section 2.

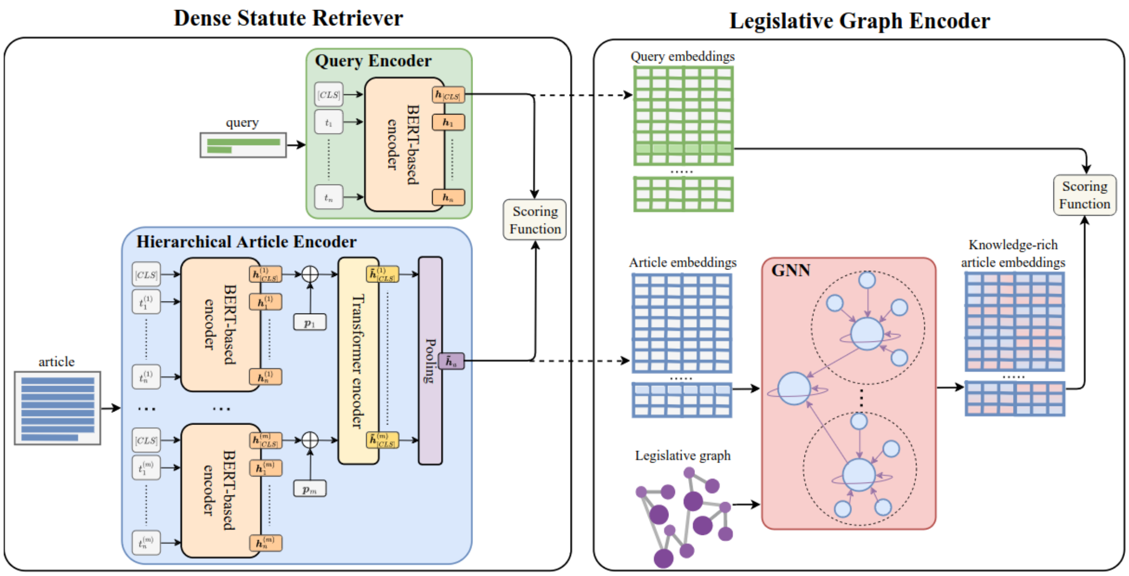
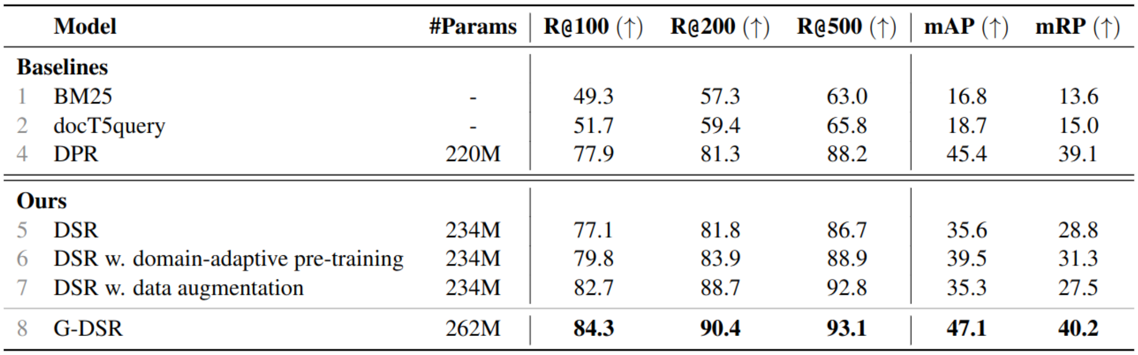


Figure 10: Graph-augmented Dense Statue Retriever (G-DSR) (Louis et al., 2023)

labeled by experienced jurists with relevant articles from a corpus of 22,600+ Belgian law articles" they had created for another paper Louis and Spanakis (2022). (https://github.com/maastrichtlawtech/bsard. Results are in Figure 11.

The authors note that compared to standard retrieval models (BM25, docT5query, and DPR), by 30% on recall@*k* (macro averaged recall at cutoff *k*) and 25% on mAP and mRP (mean average precision and mean r-precision). Further, while the use of the LGE did not dramatically change the rank-unaware results, it did improve rank-aware performance by 12%. All this indicates that G-DSR may be well worth investigating for the regulatory-related document use case.



	Model	#Params	R@100 (↑)	R@200 (↑)	R@500 (↑)	mAP (↑)	mRP (↑)
Baselines							
1	BM25	-	49.3	57.3	63.0	16.8	13.6
2	docT5query	-	51.7	59.4	65.8	18.7	15.0
4	DPR	220M	77.9	81.3	88.2	45.4	39.1
Ours							
5	DSR	234M	77.1	81.8	86.7	35.6	28.8
6	DSR w. domain-adaptive pre-training	234M	79.8	83.9	88.9	39.5	31.3
7	DSR w. data augmentation	234M	82.7	88.7	92.8	35.3	27.5
8	G-DSR	262M	**84.3**	**90.4**	**93.1**	**47.1**	**40.2**

Figure 11: Paper 4 precision and recall on BSARD data (Louis et al., 2023)

7.2 Analysis: Why this may not work for regulatory-related documents

While an *F1* score was not given, we can calculate one from the provided macro averaged precision of 47.1, and the R@100 of 84.3. I selected R@100 versus R@200 or R@500 because 100 search return results is already a lot to have users sort through looking for the correct answers.

$$F1 = \frac{2 \times mAP \times r@100}{mAP + r@100} = 60.4$$

This is not great. Although the high recall in Figure 11 is a positive for the regulatory-related document use case because, as noted earlier, in the corporate use case recall is more important than accuracy to ensure nothing has been missed.

Another point not taken into account by the LGE is that regulations cross link beyond the straightforward hierarchy as depicted in Figure 3 where 12 CFR 21.11 links to 12 USC 1813, which is in an entirely different document tree (regulations versus statutory law). Moreover, 12 CFR 21.11 also links to the Bank Secrecy Act (31 USC 5311), but contains only the name of the Act and not the direct citation. The authors acknowledge this as a an area for future research.

> *Therefore we believe that considering richer legal graph structures, especially legal citation networks could increase effectiveness even more. However, building such citation networks from raw text requires considerable text-processing effort.* (Louis et al., 2023)

While this is not necessarily a reason why this approach will not work for regulatory-related documents, it does pose both an opportunity and a challenge to determine how to find the citation cross-references and to manage the much larger and more complex graph that will no longer be a tree and likely include cycles.

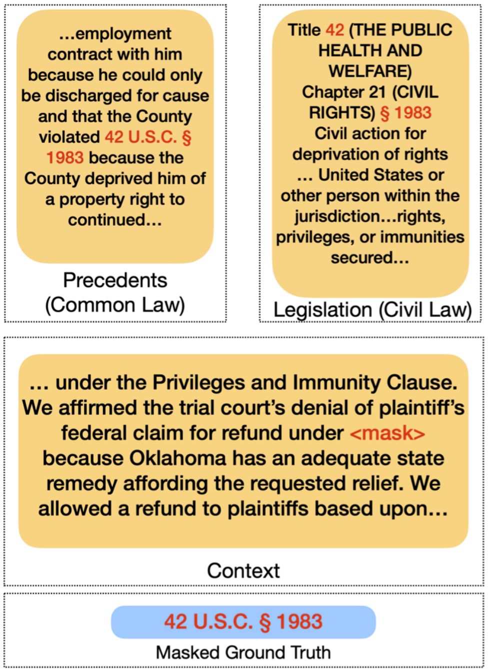
8 Paper 5: Prototype-Based Interpretability for Legal Citation Prediction (Luo et al., 2023)

My fifth paper is Luo et al. (2023), in which the authors aim to identify the most relevant legal citation for a text passage, which they call *legal citation prediction* (LCP). Their model uses *prototypes*, which are representative training examples with masked input for each target label (i.e., citation) and are shown in Figure 12. This figure also shows the overall use case. Prototypes are constructed from:

- **Context**, the input training text and surrounding context (C) tokens. As shown in Figure 14, $C \in [0, \pm 2, \pm 4]$ sentences before and after the input sentence.
- **Precedents** (Preced), other text passages with the same target label (citation).
- **Provisions** (Provis), text of laws (common law and civil law) that have been labeled with their citations.

This architecture is illustrated in Figure 13 where provisions, precedents, and input context (up to ± 4 sentences) are embedded with a RobERTa-based encoder and then measured for similarity using L2 norm. The model was trained using US Federal court cases from the PACER dataset `https://pacer.uscourts.gov` (preced), and the content of related US Code (provis). These were then scanned with regex patterns to provide gold labels, resulting in 175,741 labeled and linked documents.

The training loss function is below. $f(x)$ is the embedding of the input, with y as the expected label (citation). S is the L2 normed similarity score between the prototypes (preced and provis



...employment contract with him because he could only be discharged for cause and that the County violated 42 U.S.C. § 1983 because the County deprived him of a property right to continued...

Title 42 (THE PUBLIC HEALTH AND WELFARE) Chapter 21 (CIVIL RIGHTS) § 1983 Civil action for deprivation of rights ... United States or other person within the jurisdiction...rights, privileges, or immunities secured...

... under the Privileges and Immunity Clause. We affirmed the trial court's denial of plaintiff's federal claim for refund under <mask> because Oklahoma has an adequate state remedy affording the requested relief. We allowed a refund to plaintiffs based upon...

Figure 12: Prototype with provisions and precedent (in this figure, context) and training approach on same with masking (Luo et al., 2023)

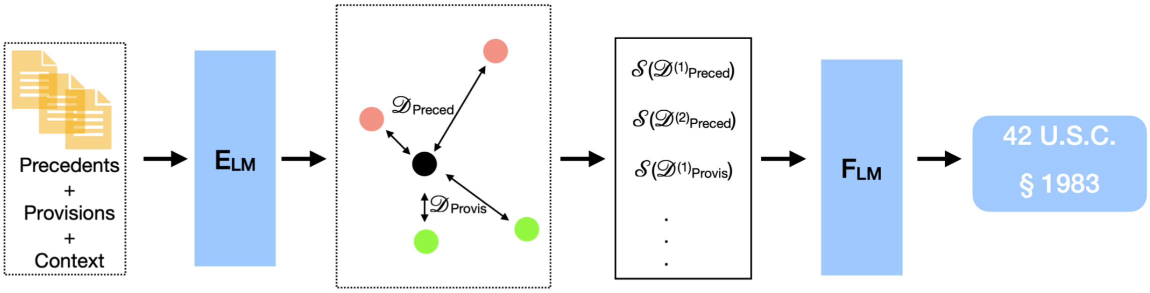


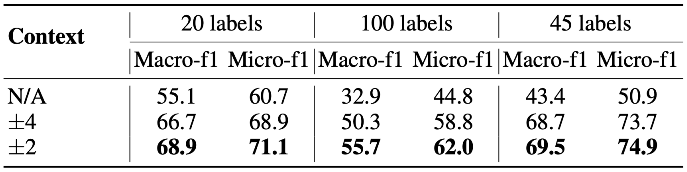
Figure 13: Prototype model with precedents, provisions and input text (in this figure, context). E_{LM} is a RoBERTa-based encoder. The black dot is the embedding of the input text with surrounding context (up to ± 4 sentences). The pink dots are embeddings of precedents (masked training data), and green dots are embeddings of provisions (law). D is the L2 norm distance between prototypes and input embeddings. F_{LM} is a feedforward layer to obtain top scoring provision. (Luo et al., 2023)

combined) and training data. *BCELoss* is standard binary cross-entropy loss. And learned hyperparameters λ and δ for weighting *preded* and *provis*, respectively.

$$L = \frac{1}{n}\sum_{i=1}^{n} BCELoss(c \circ S \circ f(x_i), y_i) + \lambda D_{preced} + \delta D_{provis}$$

8.1 Analysis: Why this may work for regulatory-related documents

As with paper 4 (Louis et al., 2023), this paper directly seeks to solve a problem closely related to the regulatory-related documented use case. Moreover, the *preced* approach could likely be directly mapped to corporate documents and making use of their references to regulations—either directly or inferrable through links such as procedures related to a policy that cites a regulation. Results in Figure 14 also appear initially promising, with an F1 as high as 74.9.



Context	20 labels		100 labels		45 labels	
	Macro-f1	Micro-f1	Macro-f1	Micro-f1	Macro-f1	Micro-f1
N/A	55.1	60.7	32.9	44.8	43.4	50.9
±4	66.7	68.9	50.3	58.8	68.7	73.7
±2	**68.9**	**71.1**	**55.7**	**62.0**	**69.5**	**74.9**

Figure 14: Paper 5 F1 (Louis et al., 2023)

8.2 Analysis: Why this may not work for regulatory-related documents

Unfortunately, on deeper examination of the results, all is not so optimistic. The highest F1 scores are for 20 and 45 labels, and the lowest, 55.7 macro-F1, for 100 labels. This means that the model does worse when the potential number of citations to select from is larger. And for the regulatory-related document use case, the number of regulatory citations that a typical financial institution must comply with is in the multiple 1000s.

Also apparently disappointing is the feedback received from three legal expert volunteers. They noted many of the citations were procedural—such as how to file a lawsuit—rather than useful for legal arguments. The experts conclude, “if it is easy to predict a citation from keywords, the task would be trivial for a lawyer.”

For the regulatory-related documents use case, however, this may not be a concern. It may even be advantageous, as labeling corporate documents with appropriate, direct and “obvious” citations, is the use case. Also, avoiding any use of expensive lawyer time, no matter how trivial, would be beneficial.

9 Conclusion

It is, of course, not fair to compare these papers numerically as they used different datasets, evaluation mechanisms, and approaches. Nevertheless, Table 1 contains selected statistical results from the papers that—with a large grain of salt—represent the authors’ views of the best that their given models could achieve. To be blunt, there is clearly a long way to go to be useful for the purposes of the papers, much less the regulatory-related document use case.

One possible overarching reason for this is explained by a chart in paper 5 (Luo et al., 2023) of legal citations versus how often they are referenced, copied here as Figure 15. The long-tail/Zipf’s Law is very much applicable. Most citations occur rarely and a handful dominate, leading to overfitting and weak results for most citations. This appears true for the cross-document co-reference papers. I strongly suspect, this is also true in corporate documents for the regulatory-related document use case.

Nevertheless, there is reason for hope. While the papers did have different ideas, there is a commonality of looking for ways to combine semantic knowledge—especially legal domain specific knowledge—with statistical or neural models. In this vein, paper 4 (Louis et al., 2023) achieved an impressive 84 on recall with their combination of a statutory structure graph and BERT-based embeddings. I believe there is more work to be done, but there is an opportunity to solve a critical problem for regulated businesses, and these papers provide an excellent foundation for future research.

Paper	Citation	Precision	Recall	F1	Comments
1	(Dutta and Weikum, 2015)	60	65	62	B^3 on NY Times
2	(Ravenscroft et al., 2021)	50	69	58	MUC using CA-V model
3	(Shaffer and Mayhew, 2019)	77	56	65	using neural model (stripped)
4	(Louis and Spanakis, 2022)	47	**84**	60	F1 estimated from provided data
5	(Luo et al., 2023)	-	-	62	Micro-F1, 100 labels

Table 1: Selected precision, recall, and $F1$ (rounded)

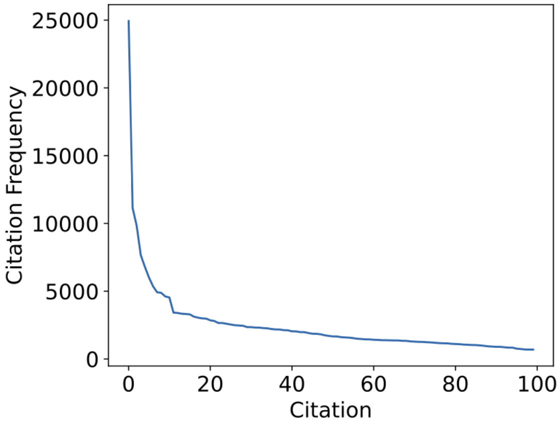


Figure 15: Long tail of citations (Luo et al., 2023)

10 Afterword

My reading of these papers has encouraged me to try and do some analysis myself. I have started a GitHub repository `https://github.com/dsagman/CFR-Parsing`. I have written a scraper to download the HTML contents of the US Code of Federal Regulations and begun initial examination of cross-references within the data. The goal being to create a training set of labeled data relating paragraphs of the CFR to each other via their existing *href* tags. Given that there are also links to other text, such as the US Code and Federal Register, there is the possibility to expand the labeled data set further.

From there it would be a next step to build a neural model to try and predict related references using off-the-shelf models or to try and adapt one or more of the models in this paper.

This is all for the future, and for now, I present an initial examination of the cross reference links within chapter 1 of title 12 of the CFR in Figure 16. Code is available in the above GitHub repository.

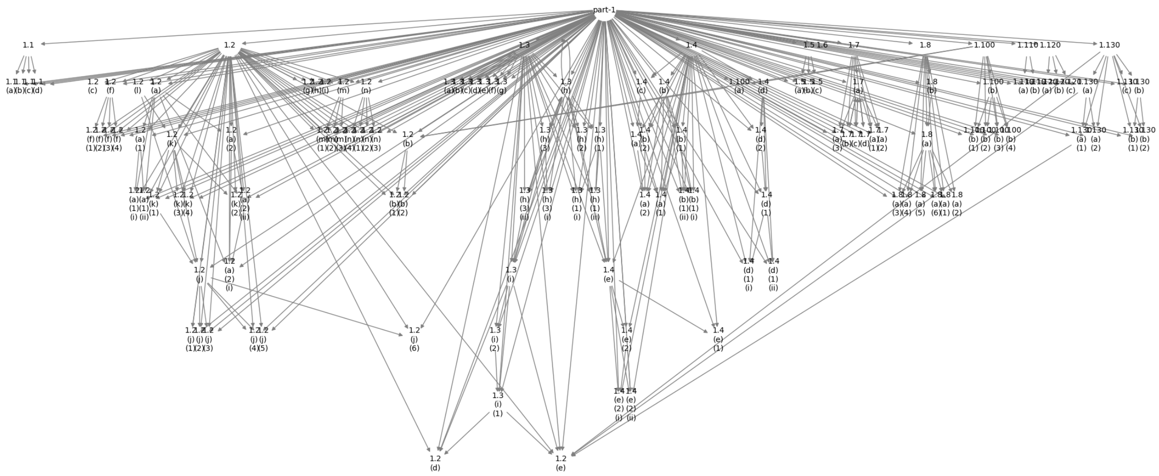


Figure 16: Cross references with 12 CFR 1

References

Sourav Dutta and Gerhard Weikum. 2015. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. *Transactions of the Association for Computational Linguistics*, 3:15–28.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. Finding the law: Enhancing statutory article retrieval via graph neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2761–2776, Dubrovnik, Croatia. Association for Computational Linguistics.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in french. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6789–6803, Dublin, Ireland. Association for Computational Linguistics.

Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Prototype-based interpretability for legal citation prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4883–4898, Toronto, Canada. Association for Computational Linguistics.

James Ravenscroft, Amanda Clare, Arie Cattan, Ido Dagan, and Maria Liakata. 2021. CD^2CR: Co-reference resolution across documents and domains. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 270–280, Online. Association for Computational Linguistics.

Robert Shaffer and Stephen Mayhew. 2019. Legal linking: Citation resolution and suggestion in constitutional law. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.