LegalDiscourse: Interpreting When Laws Apply and To Whom

Anonymous ACL submission

Abstract

While legal AI has made strides in recent years, it still struggles with basic legal concepts: *when* does a law apply? *Who* it applies to? *What* does it do? We take a *discourse* approach to addressing these problems and introduce a novel taxonomy for span-and-relation parsing of legal texts. We create a dataset, *LegalDiscourse* of 602 state-level law paragraphs consisting of $3,715$ discourse spans and $1,671$ relations. Our trained annotators have an agreement-rate $\kappa > .8$, yet few-shot GPT3.5 performs poorly at span identification and relation classification. Although fine-tuning improves performance, GPT3.5 still lags far below human level. We demonstrate the usefulness of our schema by creating a web application with journalists. We collect over $100,000$ laws for 52 U.S. states and territories using 20 scrapers we built, and apply our trained models to $6,000$ laws using U.S. Census population numbers. We describe two journalistic outputs stemming from this application: (1) an investigation into the increase in liquor licenses following population growth and (2) a decrease in applicable laws under different under-count projections.

1 Introduction

AI practitioners have long explored how to analyze legal documents – i.e. laws, court opinions and regulations (Mehl, 1958). Automatic legal question answering[1], document generation[2], and motion-filing (Gibbs, 2016) are in commercial use (Dale, 2019) and the legal reasoning capabilities of next generation of NLP models are being assessed (Guha et al., 2023), including by testing against the bar exam (Katz et al., 2023).

However, as noted by Dehio et al. (2022), GPT3 models fail when confronted with simple, yet ambiguous conditions present in legal rules (Bommasani et al., 2021), a challenge well-documented in earlier

...in counties having a metropolitan form of government and in counties having a population of not less than three hundred thirty-five thousand (335,000) nor more than three hundred thirty-six thousand (336,000), according to the 1990 federal census or any subsequent federal census, the magistrate or magistrates shall be selected and appointed by and serve at the pleasure of the trial court judge...

Figure 1: Paragraph from a sample law, Tennessee § 36-5-402. The colored blocks represent the following legal discourse elements from our schema: PROBE, TEST, SUBJECT, CONSEQUENCE, OBJECT (see Section 2). We train LLMs to identify these spans and build a web application to aggregate these span tags across state-level laws.

Transformer-based models like BERT as well (Zhong et al., 2020; Holzenberger et al., 2020). Additionally, the majority of legal study has been focused a few domains, like contracts (Koreeda and Manning, 2021; Hendrycks et al., 2021), privacy policy (Wilson et al., 2016; Zimmeck et al., 2019), and corporate law (Wang et al., 2023), and the kinds of tasks heretofore studied have been highly domain specific[3]. Other domains, such as state-level administrative law, remain relatively understudied, despite their importance to policy makers, journalists and academics.

We see the need to introduce a unified mode of study that can quickly incorporate new areas of law. In this work, *we develop a uniform discourse schema for characterising a legal text*. *Discourse analyses*, or the study of functional role of text and its relations within in a document (Carlson et al., 2003; Prasad et al., 2008), has been successfully applied to areas like argumentation (Eckle-Kohler et al., 2015), dialogue (Chen and Yang, 2021) and journalism (Spangher et al., 2022, 2021). In journalism, for instance, Choubey et al. (2020) use a unified discourse schema to describe textual relations between diffuse domains of journalism.

In this work, we develop a *legal discourse* schema to address this need, which we apply to state-level legal

[1] https://www.chatbotsecommerce.com/nrf-launches-parker-first-australian-privacy-law-chatbot/

[2] https://legal.thomsonreuters.com.au/products/contract-express/, https://turbotax.intuit.com/

[3] An example of a domain-specific task: *"Classify if the clause limits the ability of a party to transfer the license being granted to a third party"* from Hendrycks et al. (2021).

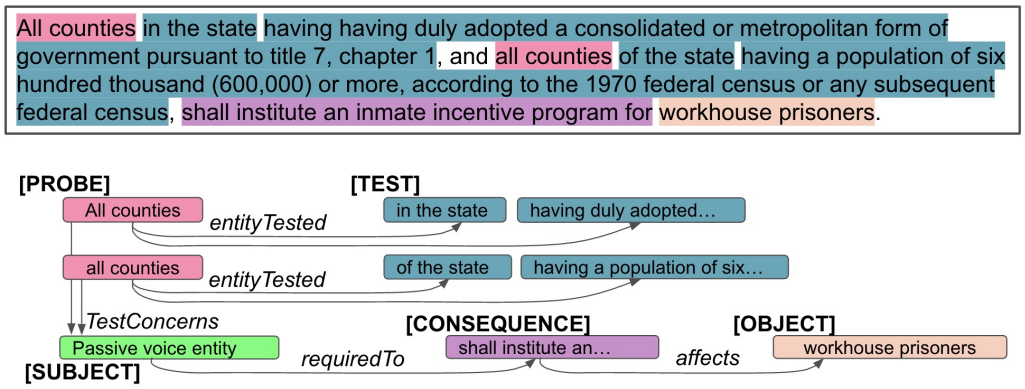

Figure 2: A sample span-and-relation discourse tree generated from a paragraph of legal text. Above, the highlighted text shows the original law text with discourse-spans annotated. Below, relations are drawn between discourse blocks, shown with double-black curved lines and categorically annotated. Note that the SUBJECT responsible for carrying out the CONSEQUENCE is passively implied.

texts. At the core, our schema seeks to answer the following key questions: (1) When does this law apply? (2) What are the consequences? (3) Who is affected? We show that LLMs struggle to model this schema, yet it is useful for human practitioners. In sum, our contributions are:

In sum, this paper makes three key contributions:

1. **A Legal Discourse Schema**: We develop a legal discourse schema, consisting of 8 span level and 21 relational classes, some of which are shown in Figure 1 and 2. We annotate 602 state-level laws, with 3,715 spans and 1,671 relations. We show that our schema can be labeled with high inter-annotator agreement, yet fine-tuned LLMs struggle.

2. **Web Scraping Public Domain U.S. State Law**: We scrape over 100,000 law documents from 52 U.S. states and territories with 20 web-scrapers we build. These scrapers are robustly designed to overcome attempts to block scraping.

3. **Searching and consuming model output**: We show the practical impact of our work by presenting a web-app we built to help users navigate our dataset. We present two outputs produced by journalists using our interface to study 6,000 laws involving 2020 U.S. Census counts.

We outline our discourse schema and modeling in Section 2. We next discuss our dataset collection process, including the web-scrapers we release for gathering public-domain U.S. state law text (Section 3.1). In Section 3.2 we describe our lightweight and modular span and relation annotation interface which we used to collect data. Next, in Section 6, we describe our web-app, where we surface our model's output to journalists and engage volunteers to improve our annotations. Finally, we discuss an ongoing use-case to illustrate how one might use our app in Section 6.2.

2 A Legal Discourse Schema

A legal rule is a *hypothetical imperative* (Engisch et al., 2018), or a conditional consequence. Reasoning about these rules requires practitioners to understand how and whether conditions of the law are met; what the consequences are (Dehio et al., 2022); and *who* is affected by these consequences.

As shown in Figure 2, modeling the different components of a legal doctrine as *discourse units* and how they interact as *relations* can be an effective way of discern meaning (Carlson et al., 2003; Prasad et al., 2008). Identifying these parts poses a basic test of a model's legal reasoning and can also lead to practical use-cases (as Spangher et al. (2022) showed in the journalism domain). We introduce the key parts in our schema, starting with span annotations and then relations.

2.1 Span-Level Schema

The 8 discourse elements we identify in our schema are SUBJECT, OBJECT, PROBE, CONSEQUENCE, TEST, EXCEPTION, DEFINITION and CLASS. The first three elements are nearly always be entities (noun phrases), and the rest are predicates (verb phrases) or prepositional phrases.

The first three elements of our schema, SUBJECT, CONSEQUENCE, and OBJECT, capture how law dictates first-degree interactions between entities, inspired by seminal work done by Gardner (1984). We describe each in turn.

- A SUBJECT is an entity that gains powers or restrictions under a law. (e.g. *"The **trial court judge** shall adjudicate property disputes between claimants."*) Subjects aren't always explicit, and can be expressed passively[4].

[4]Example of a Passive SUBJECT (and passive OBJECT) : *Taxes shall be collected at the beginning of every month.* The SUBJECT and OBJECT, "Tax-collector" and "Tax-payer", are not actively expressed.

- The CONSEQUENCE is the specific power or restriction conferred by the law. Consequences nearly always are attributed to the subject, either passively or explicitly. (e.g. *"The trial court judge **shall adjudicate property disputes** between claimants."*)

- An OBJECT is an entity (noun phrase) affected by the subject, under a law. Typically, when the subject gains powers, the object usually faces more restrictions; if the subject faces restrictions, the object usually faces fewer restrictions. (e.g. *"The trial court judge shall adjudicate property disputes between **claimants.**"*) Like subjects, objects are not always present in the text, or might be expressed passively[5].

Sometimes, the SUBJECT-CONSEQUENCE-OBJECT involves a longer chain than a 1-hop relationship[6]. In these cases, a entity is both an OBJECT and a SUBJECT. We label this entity as an OBJECT to prioritize the first CONSEQUENCE.

The next three elements in our schema, TEST, PROBE and EXCEPTION, indicate when laws apply.

- A TEST is an explicit condition applied to an entity (i.e. an OBJECT, SUBJECT or PROBE) that determines *when* a SUBJECT-CONSEQUENCE-OBJECT relation holds. (e.g. *"In counties **with a population above 10,000**, the trial court judge shall adjudicate... unless claimants settle."*

- A PROBE is an entity to which a TEST is applied to that is *not* a SUBJECT or an OBJECT. If the TEST is applied to a SUBJECT or an OBJECT, there may not be a need for a PROBE. *"In **counties** with a population above 10,000, the trial court judge shall adjudicate... unless claimants settle."*

- An EXCEPTION is a corollary to a TEST; it specifies when a law does NOT apply. An exception usually modifies a TEST *"In counties with a population above 10,000, the trial court judge shall adjudicate... **unless claimants settle**."*.

Finally, the remaining two classes in our schema, DEFINITION and CLASS, serve to more fully characterize the entities mentioned in legal text. These terms have already been well-described in the literature (Tobia, 2020; Dehio et al., 2022) and incorporated into tasks (Guha et al., 2023). We give definitions in Appendix B. For examples of all span-level discourse types, see Appendix A, Table 6.

[5]Example of a SUBJECT-CONSEQUENCE relation without an OBJECT: *The trial court judge shall begin session at or before 9am.*

[6]Example of a SUBJECT-CONSEQUENCE-OBJECT that is greater than 1-hop: "The magistrate shall designate to the county clerk, who shall adjudicate among taxpayers"

2.2 Relational Schema

We define 21 relational categories during our annotation process. There are two categories of relations. (1) The first category occurs between discourse units of *different types*. The type of these relations is usually singular based on the type of the discourse units (e.g. a TEST-PROBE relation means that the TEST is being applied to the PROBE entity), so we do not enumerate them here (we give full definitions in Appendix B). (2) The second category applies between discourse units of the *same type*. These are typically simple grammatical and logical relations. For instance, **sameEntity** indicates that two entities are instances of the same class of entity or the same instance of an entity. **Or**, **And** refers to how two predicate interact (e.g. if test_1 *OR* test_2 is passed...).

3 Dataset Creation

In this section, we describe how we operationalized the schema discussed in Section 2. We scrape a dataset of all state-level laws from 18 states in the U.S., which we discuss in Section 3.1. We then sample a set of paragraphs to annotate. We build an annotation framework, described in Section 3.2, and enlist four annotators, who collectively annotate 602 law paragraphs.

3.1 Dataset Construction

Our full legal dataset comprises the more than 100,000 active state-level laws in the United States. We compile this dataset by building a scraper for a public-domain law website called Justia.[7] We then manually audit the output collected by Justia by comparing to state websites and find 19 states where either Justia is incomplete, not updated, or unparsable.[8] We build individual state-level parsers for these states.

State law is public domain,[9] yet it is often inaccessible for bulk downloads and web scraping. For instance, many websites license LexisNexis, a for-profit company, as the official provider for their state codes[10]. Although these websites are publicly accessible, they employ a range of mechanisms (e.g. timeouts, dynamically-generated URLs, cookie-based access) that make them difficult to scrape.[11] To circumvent these, our scrapers are robust and mimic human web-browsing behavior. We develop a generalized scraper for Lexis-

[7]https://www.justia.com/

[8]Some of the laws provided by Justia, such as those for Colorado, contain data in PDF files (see https://law.justia.com/codes/colorado/2019/), which, due to formatting, have a high OCR error rate, so in these cases we we extract directly in these cases.

[9]https://fairuse.stanford.edu/overview/public-domain/welcome/

[10]Ex. Colorado, Georgia and Tennessee: http://www.lexisnexis.com/hottopics/colorado, http://www.lexisnexis.com/hottopics/gacode, http://www.lexisnexis.com/hottopics/tncode

[11]The practical effect of mechanisms to block bulk downloads is the hindrance of law corpora collection for journalistic or academic study.

	% annots	% of docs	# / doc
TEST	28%	91%	2.4
SUBJECT	20%	95%	1.7
CONS.	19%	83%	1.8
OBJECT	15%	69%	1.7
PROBE	9%	46%	1.5
CLASS	6%	34%	1.5
DEF.	2%	11%	1.6
EXC.	1%	6%	1.1

Table 1: The prevalence of different discourse units across our annotated dataset. The left column shows the percentage of units across all annotations. Center shows the percentage of documents in our corpus that have at least one discourse unit. Right shows the average number of units per document, when present.

Relation	Percentage
ENTITY ↔ PREDICATE	61%
ENTITY ↔ ENTITY	20%
PREDICATE ↔ PREDICATE	19%

Table 2: Types of relations common in our corpus. ENTITY includes: SUBJECT, OBJECT and PROBE discourse units. PREDICATE includes all others.

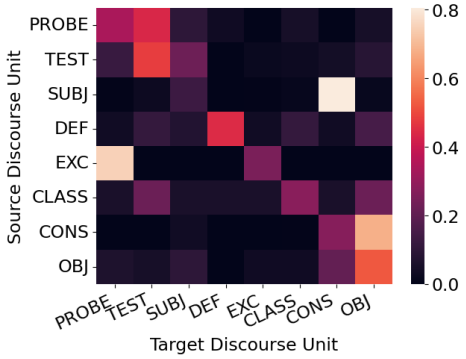

Figure 3: The conditional likelihood of a target discourse class, given a source discourse class. In other words, the color scale is $p(t|s)$ where s is the source node and t is the target node.

Nexis Public Access websites using `scrapy`[12] and `selenium-webdriver`[13]. In order to scrape Justia, we launch three Google Compute Engine (GCE) instances for a total of 60 compute hours[14].

3.2 Annotation

We recruited 4 annotators, including one former journalist and 2 undergraduate researchers[15]. We trained all of the annotators until they were achieving above an 80% accuracy in both span and relation identification tasks, based on a gold-label set that we constructed. After reaching this agreement level, we begin accepting completed tasks from annotators. Together, the annotators annotated 602 laws, with a 10% overlap, from which we calculated a $\kappa = .8$

We built a Javascript-based framework to handle span and relation tagging and (1) serve as a standalone webapp for annotators (2) compile to Amazon Mechanical Turk (AMT) tasks[16] (3) integrate into a web-site built for journalists using our work (described in Section 6). Although many NLP-focused annotation tools exist[17] we found that none were flexible enough to be integrated easily into larger websites or automatically generate AMT tasks.[18] We plan to distribute our interface as a stand-alone Javascript package. For more details about the annotation interface, see Appendix C.

3.3 Dataset Statistics

Corpus Description The length of the legal paragraphs we annotate averages 490 characters. The types of content that we focused on in our sample included topics on Government, education and environment. Certain states in our sample emphasized different topics. For example, California has a higher proportion of laws aimed at Poverty and Development compared with Tennessee, which has a higher proportion of laws focused on Administration (see Appendix A for more information and visualizations).

Discourse-level Analysis Discourse unit-level statistics vary widely. As can be seen in Table 1, TEST and SUBJECT are the most common discourse unit, accounting for 48% of all span-level annotations. TEST occurs in 91% documents. Surprisingly, EXCEPTION units were relatively rare, accounting for only 1% of annotations and occurring in only 6% of documents. There are many more TEST units per document, at 2.4 TEST units, than other elements.

Relation-level Analysis Next, we analyze the nature of the relations between discourse units. Two discourse spans are much more likely to directly relate if they are closer together in the law text. 62 characters, on average, separate discourse units with relations, while 195 characters, on average, separate all pairs of discourse units without relations. In Section 4.3, we describe how we

[12] https://scrapy.org/

[13] https://www.selenium.dev/.

[14] We will release our code for scraping with Docker images created to perform these scrapes. Given the difficulty in creating this dataset, we believe these routines constitute a considerable resource for academic inquiries into state-level law.

[15] We compensated the undergraduate researchers fairly at a rate of $20 per hour through AMT, according to University policy

[16] https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference_HTMLQuestionArticle.html.

[17] There were 87 frameworks as of Neves and Ševa (2021)'s count, including BRAT (Stenetorp et al., 2012), YEDDA (Yang et al., 2017) and WebAnnon (Yimam et al., 2013)

[18] We will release the annotation code as part of this framework

balance our training datasets to remove this adjacency bias.

Figure 3 shows the likelihood of transitioning to a target discourse type, given a source discourse type. We order the x and y axes by the most likely starting points of discourse elements in a document (Note, in Figure 6, that discourse elements that are appear first in the document to be connected with discourse elements later. See Table 5 in Appendix A for more information). We see a strong diagonal bias: all discourse elements are likely to transition to elements of the same type. We also notice the strong SUBJECT $\rightarrow$ CONSEQUENCE and CONSEQUENCE $\rightarrow$ OBJECT relation, as well as the PROBE $\rightarrow$ TEST relation. This reinforces insights by (Gardner, 1984), (Engisch et al., 2018) and (Dehio et al., 2022) about the key role of *hypothetical imperative* language in legal texts (discussed in Section 2).

On the other hand, we find that several categories of relation are simply unlikely to ever occur. For instance, EXCEPTION is almost never applied to CONSEQUENCE. We hope in future work to investigate if these patterns hold up across a wider body of legal text. See Appendix A for more details. We test the implication of this in Secion 4.3.

4 Legal Entity and Relational Modeling

We frame a new task using the data we collect: Legal Entity and Relational Modeling, or *extracting legally significant spans and their relations*. This task is analogous to end-to-end relation extraction (ERE) (Kameyama, 1997). We will first describe two subtasks that traditionally compose ERE, and how legal discourse can be modeled in this framework, then we will discuss methods, with a particular focus on how we can use this setup to interrogate the reasoning capabilities of large language models.

4.1 Tasks and Datasets

Span-Level Tagging Given a document X of n tokens $x_1, ..., x_n$, let $S = \{s_1, s_2 ... s_m\}$ be all possible spans in X. Let ζ be a set of predefined span-types, in our case we use a subset of our discourse tags: SUBJECT, CONSEQUENCE, OBJECT, TEST, PROBE and EXCEPTION. We focus on these types because they have within-text consequence, compared with, DEFINITION and CLASS, which are primarily about adding context and helping to reason across texts (Buey et al., 2016). Our goal, then, is to predict an entity type $y_e(s_i) \in \{\zeta, \epsilon\}$, where ϵ is the null class. In legal reasoning, this subtask can help test a model's awareness of the function of each span of text.

We filter our task dataset so that each document has at minimum two of the primary 6 spans, and we additionally remove spans that are at most one word, as these were the most ambiguous for our annotators to agree on. This leaves us with 3,559 spans across 413 documents. We measure classification accuracy using F1 per class, and we consider a span to be valid if it contains 80% of more of the same words, after removing stop words and punctuation, and being no longer than twice in length as the annotated example.

Relation Extraction Let R be a set of pre-defined relation types. We seek, for every pair of spans $s_i \in S, s_j \in S$ to predict a relation-type, $y_r(s_i, s_j) \in \{R, \epsilon\}$, where ϵ is the null class. We consider two versions of this task: *detection* and *classification*. Detection involves simply predicting $y_r(s_i, s_j) \in \{I, \epsilon\}$, where I indicates there exists any relation $\in R$, and classification is the classical relation-type detection. This task can help test a legal model's ability to identify which spans are modified by a given span.

To construct a challenging legal relation classification dataset, we take a subset of relations $\hat{R} \in R$ that are observed occurring between span pairs of different span-types. In other words, we take relations $r \in \hat{R}$ where $|\{y_e(s_i), y_e(s_j)\}| > 1 \; \forall_{i,j} \; s.t. \; y_r(s_i, s_j) = r$. This allows us to focus less on modeling the semantics of each span's type and more on the relation between them. We additionally sample negative examples, i.e. $y_e(s_i, s_j) = \epsilon$.

Finally, we notice that discourse units that are more proximal in the text are more likely to be related, as noted in Section 3.3. We find in early trials that our models were overfitting to proximity in text and not generalizing well to cases where relations are more distant. So, to make the task more challenging, we sample negative examples that the same distribution of offsets our labeled examples. We are left with 1,482 datapoints. We measure model accuracy using F1, focusing on three main groupings: relations between entities and entities (ENT$\leftrightarrow$ENT), relations between entities and predicates (ENT $\leftrightarrow$ PRED) and relations between predicates and predicates (PRED $\leftrightarrow$ PRED).

4.2 Baselines

Relation extraction is a widely studied field, with classical and current work focusing on modeling each subtask separately (Sang and De Meulder, 2003; Zelenko et al., 2003), as well as end-to-end modeling (Li and Ji, 2014). As such, we build upon two recent methods focused on each approach:

- PURE (Zhong and Chen, 2020): separately models two different embedding spaces, one focused on span identification and the other focused on relation extraction, using masked language modeling (Devlin et al., 2018).
- ASP (Liu et al., 2022): trains a generative T5 model (Raffel et al., 2020) to create structured predictions.

4.3 Generative Modeling

Recent work has shown that large language models can also be effective relation predictors (Wan et al., 2023). To test this hypothesis, and to add to a growing body of work focused on benchmarking LLMs for legal tasks (Guha et al., 2023), we format our tasks as generation

	SUBJECT	CONS	OBJECT	PROBE	TEST	EXC	Macro	Micro
Baselines								
ASP (Liu et al., 2022)	35.7	39.4	26.3	38.9	44.6	33.3	37.7	36.6
PURE (Zhong et al., 2020)	41.5	45.2	25.0	56.1	17.3	36.4	34.3	36.5
GPT3.5								
0-shot	34.4	9.7	14.8	13.4	35.4	54.7	27.1	22.7
3-shot	31.7	23.3	20.4	28.2	43.9	46.2	32.3	30.1
5-shot	30.7	24.1	15.9	30.8	49.8	45.2	32.8	30.8
8-shot	29.7	23.4	15.8	33.5	48.4	53.8	34.1	31.0
GPT fine-tuned	42.1	49.9	35.9	34.9	53.0	56.0	45.3	44.3

Table 3: F1 scores shown for span-identification for our 6 primary discourse elements: SUBJECT, CONSEQUENCE, OBJECT, PROBE, TEST and EXCEPTION. Average Precision, Recall and F1 across all samples are shown. Although fine-tuning improves performance across most categories, leading to +10-point increases in macro and micro f1-scores, although some, like EXCEPTION, are able to be handled relatively well even in zero-shot settings. F1 scores are still below human levels of agreement.

problems and fine-tune GPT3.5 models[19]. For span-prediction, we show that we can recover spans specific to discourse tags, in other words: we model model $p(s|\zeta, X)$. As each law may contain several discourse elements of each type, we ask the LLM to generate *all* elements of a certain discourse type in the body of the law (e.g. *You are a legal assistant. I will show you a paragraph of law. Which entities gains powers, restrictions or responsibilities under this law?*) See Appendix F for more examples of prompts. For our evaluation set, we sample such that 30% of data contains *no* discourse elements of the required type.

For relation *detection* we ask the model $p(I|s_1, s_2, X)$, i.e. whether a relation is present between two spans, and $p(R|s_1, s_2, X)$, what the relation is. We still allow $\epsilon \in R$, so that our experiments with GPT are comparable to the baseline models. In other words, we construct a prompt where the LLM is given the legal text and two discourse elements, and ask if they are related (e.g. *"Are span A and B related in Law X?"*). We test two different prompt settings. In the first setting, we simply give the two spans of text and the law, and ask the LLM to determine if they are related. In the second setting, we give the LLM the class labels of the discourse units, as well as definitions for what each label means. See Appendix A for examples. We test both tasks in zero-shot, few-shot, and fine-tuned settings and for each test sample, we repeatedly query the LLM for 3 trials, randomizing the few-shot examples it receives. .

5 Results and Discussion

Span-Level Tagging : Table 3 shows F1 scores from our span-tagging experiments. Interestingly, span-tagging appears to be a harder task for GPT: even after fine-tuning, GPT scores below human-level (our annotators, after conferencing and training). GPT was especially challenged by distinguishing between different entities' roles: SUBJECT, OBJECT and PROBE (GPT Fine-tuned scores 35-42 F1 on entities, compared with 50-59 F1 for predicates. EXCEPTION stands out as a particular category where even 0-shot GPT performs well.) SUBJECT and OBJECT roles can be particularly ambiguous, as mentioned in Section 2, as there are cases when an entity can be in both a SUBJECT and OBJECT role (we annotated OBJECT, in those cases). Interestingly, too, the gap between GPT and the baseline models is not as large in this task than it is in relational modeling. Perhaps our generative setup for this step, $p(s|\zeta, X)$, with 6 different prompts, allowed GPT to generate the same entity for different categories. We might see improvements by disambiguating with another model, $p(\zeta|s, X)$, when a single span is generated in multiple categories.

Our broader finding, though, is that this remains a challenging task. Although our task dataset, at 400 documents, is small relative to other language resources, the spans in our schema are syntactically low-level. The spans divide relatively well into different parts of speech, like noun phrases and verb phrases; identifying such chunks in text has long been within the capability of even classical language models (Sang and Buchholz, 2000). Future work either fine-tuning on other resources, or using law-specific models, might show improvements in these areas.

Relation Identification and Classification Table 4 show F1 scores from relation detection (Detect) and classification (Class). Relation extraction is a category where fine-tuned GPT performs just as well as our annotators. We notice, too that in some cases GPT does even better on the classification task than it does on the identification task (e.g. ENT↔PRED and ENT↔ENT). It's possible that the semantics of classification task enforce greater reasoning and justification than the identification

[19]Specifically, we use `GPT3.5-turbo` as of October 11, 2023.

	ENT ↔ PRED		ENT ↔ ENT		PRED ↔ PRED		All (Macro)		All (Micro)	
	Detect	Class.	Detect	Class.	Detect	Class.	Detect	Class.	Detect	Class.
Baselines										
ASP	26.5	14.2	4.5	3.8	4.0	2.2	13.6	6.7	19.5	11.1
PURE	73.9	64.5	15.4	5.3	45.7	38.2	49.5	40.5	63.1	53.9
GPT3.5										
zero-shot	54.9	0.0	42.5	27.1	25.2	23.2	40.8	16.8	48.5	7.2
zero-shot w. def	69.4	0.0	54.2	39.5	60.8	48.2	61.5	29.2	65.1	12.8
few-shot	50.6	55.3	56.8	53.9	40.2	34.2	49.2	47.8	50.5	51.7
few-shot w. def	72.6	60.1	68.5	65.9	65.1	35.2	68.7	53.7	70.8	56.7
GPT finetuned	82.6	85.9	76.5	88.7	81.0	65.9	80.0	80.2	81.1	82.9

Table 4: Relation Detection and Classification F1 score. We examine scores between three categories of relations: ENTITIES ↔ ENTITIES, ENTITIES ↔ PREDICATES, and PREDICATES ↔ PREDICATES. ENTITIES are SUBJECT, OBJECT and PROBE, and PREDICATES are all other discourse types. Classification is only run for discourse-type pairs where more than one relation can exist (see Section 2).

task, like in Wei et al. (2022).

The relation identification task also shows a clear different between the baseline models, which we do not observe in the span-level tagging task. One explanation for the especially poor performance of ASP (Liu et al., 2022) is that the jointly learned model requires the model to make use of more data to fully learn the embedding layers. In fact, tasks that ASP performs well on, like ACE2005 (Sang and De Meulder, 2003), have $\tilde{1}$0x more documents and annotation than our dataset. We show more details in Appendix E, Figure 7.

6 Practical Use Case: Census 2020

To get feedback on our work from a preliminary group of users, we apply our models to a domain of state-level law pertinent to journalists. In 2020, the U.S. Census count faced multiple challenges, and many researchers hypothesized that populations, especially minorities, might be inaccurately counted (Naylor, 2020; Mervis, 2019; Berry-James et al., 2020). Scant insight existed, especially on the state-level, into how population counts were being used in law[20]: the corpus of state-level laws was too large and varied for journalists to parse.

On the other hand, this provided an interesting case for discourse-based reasoning. Population counts typically get used as a relatively unambiguous TEST. Our discourse models help us identify this occurring, and then we can develop ways to parse out the specific ways population is in TEST discourse. We describe the website we built to facilitate different explorations, and then we describe two such explorations that we received permission from the journalists collaborating with us to write about. We will focus on our own contributions in these collaborations.

[20]Besides federal budgeting and Congressional representation, which have already been manually programatized (Reamer, 2018; Berry-James et al., 2020).

6.1 Website Design

We design a website to enable exploration of our dataset and modeling output. Users can perform full-text search on all laws in our database, view discourse spans schema by extracting spans across laws and correct or provide new annotations. The website's overall goal is to facilitate both ***deep*** **explorations** and ***wide*** **explorations**.

Going deep : Going "*deep*" here essentially means *subsetting the laws* first, and then analyzing discourse. The web search functionality[21] helps users do this by exploring a specific term or concept in the law's plain text or in specific discourse role (e.g. laws affecting OBJECT="taxpayer"). After the user finds samples of an interesting subsset of laws, we do a broader study of them them by using our discourse models to answer: *who* is being affect, under *what conditions*, and *how*?

Going wide : Conversly, going "*wide*" means studying discourse units and relations first, then analyzing the laws. The website includes a second functionality: allowing users can view aggregate counts of different discourse units and relations. This helps users notice patterns among the ways in which discourse was being used. After a user notices a specific pattern in discourse roles (e.g. EXCEPTION units modifying TEST units about taxes), then we can analyze the laws that include, or do not include, these elements.

In both flows, visitors can access our annotation framework, described in Section 3.2, which helped us gather more data (to be used in the future). For more on the design of our website, see Appendix D.

6.2 Case Study #1: Going Deep (Liquor Store Licenses)

We describe two example articles that are currently being explored by users of our system.

[21]Powered by ElasticSearch (Elasticsearch, 2018)

In the first example, journalists hypothesized that the allocation of new liquor licenses might be population-based. To explore this, they used the search interface; they searched for the term "alcohol OR liquor OR beverage" in the search interface and discovered that interface returned 270 laws. Together, we analyzed the breakdown of liquor-related law by state. We found that the states most likely to base liquor licenses off population counts were Tennessee, New York and Illinois. They then asked us to extract all TESTS from these laws. We found that mid-size cities would be the most likely to be impacted by a 5% or 10% undercount in population. The journalists identified key cities and are seeking sources in these areas.

6.3 Case Study #2: Going Wide (Slim Population Thresholds)

In another example, journalists explored the top-level discourse annotations. They noticed that some TESTS are based on explicit population thresholds (ex. Figure 1) and that some of these thresholds were very narrow. Working together, we compiled several keyword filters and regular expressions extract specific population thresholds. We found that, in Tennessee in particular, over 40% of all Census-related laws imposed narrow population tests of fewer than 500 people and 10% imposed tests of fewer than 100 people. This raised questions: what is the purpose of these narrowly targeted laws? Were they trying to target specific counties without mentioning them by name? The journalists are now investigating further by tracking down the authors of these laws.

7 Related Work

Although the field of AI-driven legal aids is multifaceted and growing (Kauffman and Soares, 2020), free and open-source frameworks remain few (Morris, 2019; Dale, 2019; Vergottini, 2011). Our discourse-driven web application, designed for legal exploratory analysis is one of the few AI-powered, free applications that exist, and the first to open source tools for legal document collection.

For-profit legal inquiry systems, as mentioned above, are numerous. Bloomberg Law[22], Westlaw[23], LexisNexis[24] and Wolters Kluwar[25] are the four main services for legal research (Dale, 2019), which provide subscription-based, Google-style searches. CaseText[26] and Ravel[27] were two upstart case-text search engines (although both have now been aquired); CaseText offered crowdsourced annotations and Ravel linked cases together to create visual maps of important cases (Lee et al., 2015). We similarly provide a way of collecting user-annotations, and a novel way linking together cases, although ours takes a discourse approach rather than an unsupervised clustering approach.

Various discourse schemas have been developed to understand law texts, including deontological logic-based schemas (Wyner and Peters, 2011; Zeni et al., 2015), and subject matter-specific schemas (Espejo-Garcia et al., 2019). Ours is the first discourse-based approach to take steps towards a big-data approach by setting up a framework for the ingestion of crowdsourced annotations.

Finally, outside of the legal domain, other areas have experienced a growth in academically-oriented systems for human-in-the-loop inquiry. The COVID-19 pandemic has produced a burst in NLP-driven corpora-collection (Wang et al., 2020), demonstrations (Sohrab et al., 2020; Hope et al., 2020; Spangher et al., 2020) and workshops (Verspoor et al., 2020b,a).

Such concerted effort in the NLP domain to expose resources and build open tools for subject matter experts is an inspiring guide for how NLP researchers can contribute to wider inquiries. We hope such efforts expand to other domains as well, forming a common alliance between academics, civil-minded journalists and other researchers and end-users.

8 Conclusion

We have sought to take steps towards a semantic understanding of legal texts, a goal long held in computational law (Gardner, 1984). We show that large language models, while achieving impressive results in some parts of our task, show surprisingly weak performance compared to human annotators in others. Language models have an important role to play in interpreting law and lowering the barrier of access to legal systems for citizens, journalists and academics. Our task is an important step towards assessing a sturdy foundation and opening the door to more intensive legal tasks be considered (Guha et al., 2023).

In this work, we have additionally presented three open-source components. (1) A web-app exposing a novel discourse schema and its application to state law referencing U.S. Census counts. (2) A flexible and modular annotation framework that can be seamlessly embedded into web-apps to allow visitors to contribute and update annotations. (3) A set of web-scrapers to help researchers gather public-domain legal text. We demonstrated concrete utility to facilitate journalistic exploration with our discourse schema. Our longer-term goal is to collect feedback and data, and improve our database and machine learning systems. We hope that such efforts can continue to push Legaltech (Hartung et al., 2017) into a more open and accessible domain, and make it easier to understand the laws governing our society.

[22] https://pro.bloomberglaw.com/
[23] https://www.westlaw.com/
[24] https://www.lexisnexis.com
[25] https://www.wolterskluwer.com
[26] https://casetext.com/
[27] https://home.ravellaw.com/

9 Impact Statement

There were several possible ethical considerations we encountered during this research which we wish to address.

Dataset Creation: The creation of our dataset involved scraping numerous websites, including state websites, state-licensed LexisNexis pages and https://www.justia.com. In the third case, Justia, we did not violate any terms of service. In fact, Justia's robots.txt file[28] is the most permissive possible, giving unlimited license to any crawler. It is generally accepted that robots.txt files are implied licenses of access,[29] and we did not disregard Justia's file before scraping.

Content derived from the first two categories, state law websites and official, state-licensed websites like LexisNexis are, by law, public domain (Wolfe, 2019; MacWright, 2013). Web-scraping the public domain is neither illegal nor unethical (Mehta, 2021). As we did in the body of the paper, we again emphatically criticize attempts by providers to make web-scraping difficult, and we went to lengths to overcome this.

Dataset Annotation: All parties involved in annotating our dataset received valid compensation. We relied entirely on expert researchers to collect our annotations. This included the authors of this paper. All the researchers who provided annotations for us were affiliated with our institution and compensated appropriately by our institution (we leave the determination of "appropriate" for our institution to define.)

Although we describe accommodating AMT tasks in the body of the paper, thus far, we have not used any annotations made by Turkers on AMT or by journalists/researchers using our site. If we do, we will ensure there are no ethical issues by securing university IRB approval or exemption, as deemed fit by the IRB. For the Turkers, we will calculate a payment that equals, on average, $15 an hour. For the journalists/researchers, we will have exchanged something of value (the use of our web-app) for the annotation.

Website Usage: Our website has significant accessibility limitations for the seeing-impaired and for non-English speakers. We have not addressed them in this current version, but are mindful and actively searching for options to expand accessibility.

There are two ways in which seeing-impaired users might suffer. First, blind users will not be able to read any of the site without external tools, as we have not recorded or built in any native audio-scripts, keyboard shortcuts or voice-activated commands. Besides "not containing irrelevant information" (Giraud et al., 2018), we can do more to audit our website (Tosaka, 2005) and organize the flow on our site to increase blind accessibility. Secondly, part of our website introduces users to our discourse schema by introducing them to color-coded segments of text. We are actively investigating color-schemes and other approaches that are more amenable to color-blind individuals, of which there is extensive research (Wakita and Shimamura, 2005; Jambor et al., 2021; Foti and Santucci, 2009). Because of the prototype nature of this website, we have not yet investigated these, but they are crucial next-steps.

Our website focuses on U.S.-based laws and contains only English-language text. We do not attempt, in this version, to perform translations. Our plan in the present iteration of this work was to work with U.S.-based journalists studying U.S.-based law. We have not yet undertaken a study to compare how well our discourse schema would apply to non-U.S. law, be it common or civil (Dainow, 1966). However, if this approach proves useful for journalists and researchers, we will certainly seek to undertake this.

[28] Found here https://www.justia.com/robots.txt. Such files govern the site-owners' standards for scraping and crawling.

[29] https://stackoverflow.com/questions/999056/ethics-of-robots-txt

References

RaJade M Berry-James, Susan T Gooden, and Richard Greggory Johnson III. 2020. Civil rights, social equity, and census 2020. *Public Administration Review*, 80(6):1100–1108.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

María Granados Buey, Angel Luis Garrido, Carlos Bobed, and Sergio Ilarri. 2016. The ais project: Boosting information extraction from legal documents by using ontologies. In *ICAART (2)*, pages 438–445.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Current and new directions in discourse and dialogue*, pages 85–112.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. *arXiv preprint arXiv:2104.08400*.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Joseph Dainow. 1966. The civil law and the common law: some points of comparison. *Am. J. Comp. L.*, 15:419.

Robert Dale. 2019. Law and word order: NLP in legal tech. *Natural Language Engineering*, 25(1):211–217.

Niklas Dehio, Malte Ostendorff, and Georg Rehm. 2022. Claim extraction and law matching for covid-19-related legislation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 480–490.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242.

BV Elasticsearch. 2018. Elasticsearch. *software], version*, 6(1).

Karl Engisch, Thomas Würtenberger, and Dirk Otto. 2018. *Einführung in das juristische Denken*. Kohlhammer Verlag.

Borja Espejo-Garcia, Francisco J Lopez-Pellicer, Javier Lacasta, Ramón Piedrafita Moreno, and F Javier Zarazaga-Soria. 2019. End-to-end sequence labeling via deep learning for automatic extraction of agricultural regulations. *Computers and Electronics in Agriculture*, 162:106–111.

Antonella Foti and Giuseppe Santucci. 2009. Increasing web accessibility through an assisted color specification interface for colorblind people. *IxD&A*, 5:41–48.

Anne von der Lieth Gardner. 1984. Artificial intelligence approach to legal reasoning. Technical report, Stanford Univ.

Samuel Gibbs. 2016. Chatbot lawyer overturns 160,000 parking tickets in london and new york. *The Guardian*.

Stéphanie Giraud, Pierre Thérouanne, and Dirk D Steiner. 2018. Web accessibility: Filtering redundant and irrelevant information improves website usability for blind users. *International Journal of Human-Computer Studies*, 111:23–35.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

Markus Hartung, Micha-Manuel Bues, and Gernot Halbleib. 2017. *Legal tech*. CH Beck.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257*.

Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel Weld, Marti Hearst, and Jevin West. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 135–143, Online. Association for Computational Linguistics.

Helena Jambor, Alberto Antonietti, Bradly Alicea, Tracy L Audisio, Susann Auer, Vivek Bhardwaj, Steven J Burgess, Iuliia Ferling, Małgorzata Anna Gazda, Luke H Hoeppner, et al. 2021. Creating clear and informative image-based figures for scientific publications. *PLoS biology*, 19(3):e3001161.

Megumi Kameyama. 1997. Recognizing referential links: An information extraction perspective. *arXiv preprint cmp-lg/9707009*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Marcos Eduardo Kauffman and Marcelo Negri Soares. 2020. AI in legal services: new trends in AI-enabled legal services. *Service Oriented Computing and Applications*, 14(4):223–226.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Katrina June Lee, Susan Azyndar, and Ingrid AB Mattson. 2015. A new era: Integrating today's next gen research tools ravel and casetext in the law school classroom. *Rutgers Computer & Tech. LJ*, 41:31.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412.

Tianyu Liu, Yuchen Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. *arXiv preprint arXiv:2210.14698*.

Tom MacWright. 2013. State law is public domain. what's public domain? *macwright.com*.

Lucien Mehl. 1958. *Automation in the legal world*. National Physical Laboratory.

Ivan Mehta. 2021. Us court says scraping a site without permission isn't illegal. *TNW | Security*.

Jeffrey Mervis. 2019. Census citizenship question is dropped, but challenges linger. *Science*, 365(6450):211–211.

Jason Morris. 2019. Making mischief with open-source legal tech: Radiant law.

Lorenda A Naylor. 2020. Counting an invisible class of citizens: The lgbt population and the us census. *Public Integrity*, 22(1):54–72.

Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1):146–163.

Alessandra Potrich and Emanuele Pianta. 2008. L-isa: Learning domain specific isa-relations from the web. In *LREC*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Andrew Reamer. 2018. Counting for dollars 2020: the role of the decennial census in the geographic distribution of federal funds. *Initial Analysis*, 16.

Erik F Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. *arXiv preprint cs/0009008*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Mohammad Golam Sohrab, Khoa Duong, Makoto Miwa, Goran Topić, Ikeda Masami, and Takamura Hiroya. 2020. BENNERD: A neural named entity linking system for COVID-19. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 182–188, Online. Association for Computational Linguistics.

Alexander Spangher, Jonathan May, Sz-rung Shiang, and Lingjia Deng. 2021. Multitask learning for class-imbalanced discourse classification. *arXiv preprint arXiv:2101.00389*.

Alexander Spangher, Yao Ming, Xinyu Hua, and Nanyun Peng. 2022. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6848–6866.

Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2020. Enabling low-resource transfer learning across COVID-19 corpora by combining event-extraction and co-training. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Kevin P Tobia. 2020. Testing ordinary meaning. *Harv. L. Rev.*, 134:726.

V Yasuaki Takamoto V Hideki Tosaka. 2005. Web accessibility diagnosis tools. *Fujitsu Sci. Tech. J*, 41(1):115–122.

Grant Vergottini. 2011. To go open source or not?

Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace, editors. 2020a. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online.

Karin Verspoor, Kevin Bretonnel Cohen, Mark Dredze, Emilio Ferrara, Jonathan May, Robert Munro, Cecile Paris, and Byron Wallace, editors. 2020b. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online.

Ken Wakita and Kenta Shimamura. 2005. Smartcolor: disambiguation framework for the colorblind. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 158–165.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. Maud: An expert-annotated legal nlp dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Jan Wolfe. 2019. U.s. high court to rule on scope of copyright for legal codes.

Adam Z Wyner and Wim Peters. 2011. On rule extraction from regulations. In *JURIX*, volume 11, pages 113–122.

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2017. Yedda: A lightweight collaborative text span annotation tool. *arXiv preprint arXiv:1711.03759*.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.

Nicola Zeni, Nadzeya Kiyavitskaya, Luisa Mich, James R Cordy, and John Mylopoulos. 2015. Gaiust: supporting the extraction of rights and obligations for regulatory compliance. *Requirements engineering*, 20(1):1–22.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.

Zexuan Zhong and Danqi Chen. 2020. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812*.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66.

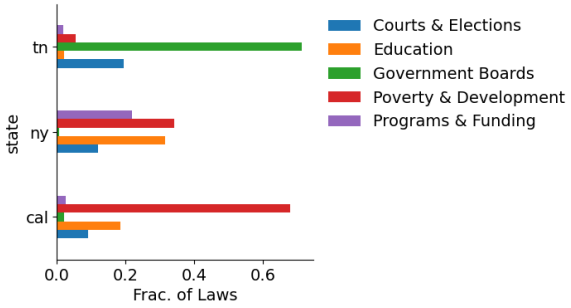

Figure 4: Topic model run over our corpus, showing 3 states. Topics are manually labeled by analyzing top words.

label	start	end
PROBE	28%	32%
TEST	36%	56%
SUBJECT	36%	47%
DEFINITION	41%	55%
EXCEPTION	42%	50%
CLASS	45%	61%
CONSEQUENCE	46%	56%
OBJECT	51%	59%

Table 5: Average start and end character positions of discourse units.

A Additional Data Analysis

We give analysis of the corpus we collected. In Table 5, we show average character positions of discourse units in the document, as a percentage of the length of the document. PROBE is most likely to occur first in a document, followed by TEST. Discourse units are less likely to occur in the second half of the document.

We examine attributes of relations between discourse elements in Figure 3 and 5. Figure 3 shows the likelihood of transitioning to a target discourse type conditioned on a source type. In Figure 2, we observe there is a strong bias for discourse elements that are appear first in the document to be connected with discourse elements later. We order the x and y axes by the most likely starting points, as given in Table 5. We see a strong diagonal bias: all discourse elements are likely to transition to elements of the same type. We also notice the strong SUBJECT $\rightarrow$ CONSEQUENCE $\rightarrow$ OBJECT relation, as well as the PROBE $\rightarrow$ TEST relation.

B Additional Schema Definitions

B.1 Span-Level Schema: Minor Classes

- A DEFINITION is a span of text serving to clarify the ordinary meaning (Tobia, 2020) of a term used in the legal text (e.g. *"**"Qualified taxpayer"** means a person or entity engaged in a trade or business within..."*)

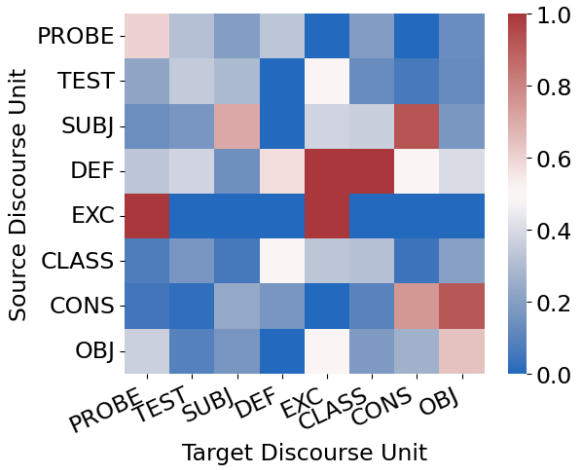

Figure 5: Likelihood of relations, normalized by the random chance of a relation occurring in a document. In other words, if n_p is the count of all annotated pairs of relations and n_r is a count of all randomly occurring pairs from a random sample of all $\binom{n}{2}$ pairs of discourse units in a law text, then the color scale is $\frac{n_p}{n_p+n_r}$. Values $> .5$ are *more* likely to be paired than random chance.

- The CLASS of an entity is a modifier that serves to disambiguate the entity from other entities. In knowledge-graph terms, CLASS specifies the source node in an *isA*-type relationship (Potrich and Pianta, 2008); specifically, the entity with a CLASS tag is a subclass of the entity without the class tag (e.g. *"The **trial court** judge shall, in..."*).

B.2 Relational Schema

We define 21 relational categories during our annotation process. The first category is of relations that occur between text-spans of different types.

- **EntityEmpoweredTo**, **EntityRequiredTo**: Indicates which SUBJECT entity (or, in rarer cases, an OBJECT entity) recieves powers or responsibilities under a CONSEQUENCE.
- **Affects**, **AffectedBy**: Indicates which OBJECT entity or entities are affected the CONSEQUENCE of the law, usually mediated through the SUBJECT.
- **TestConcerns**, **Entitytested**: Indicates which entity a TEST is applied to. This relation typically establishes conditions governing many other entities in the law, not just the entity tested; this is especially the case when the entity is a PROBE.
- **ExceptionAppliesTo**, **ConditionalTest**, **ConditionalConsequence**: Indicates which discourse unit (or, in some cases, a second TEST) is applied to. EXCEPTION and multiple levels of TEST are broadly applied to all different kinds of discourse units.
- **Comparison**: A more nebulous relational category, this forms the basis of how the conditional applicability of a law is assessed. Usually found in TEST relations, this relation-type occurs when an attribute of an entity is measured in order to make a determination about whether the conditions are satisfied and the law may be applied. This relation inspired by original attempts to support programmatic legal analysis (Gardner, 1984).)
- **EntityHasProperty**, **PropertyOf**: When any has a particular attribute or CLASS (can be used along with the **Comparison** relation and a TEST).
- **IsDefinedAs**, **DefinitionOf**: Indicates the entity being defined by a DEFINITION.

The second category of relations typically applies to spans of the same type:

- **SameEntity**: Indicates that two entities are either separate instances of the same class of entity, or they literally refer to the same instance of an entity in legal text.
- **Continuation**: Indicates that two disjointed spans of text refer to the same discourse unit. Can occur when a span is split by another discourse unit.
- **FollowedBy**: When one predicate is conducted or evaluated after the other, in logical order (e.g. in a CONSEQUENCE-CONSEQUENCE relation: *"The magistrate must attend the meetings, then they may be seated."*).
- **Or**: Either two predicates or entities are mentioned in the law, but when only one needs to be passed (in the case of a predicate), or only one entity is affected.
- **And**: Either two predicates or entities are mentioned in the law, but both need to be passed (in the case of a predicate), or both entities are affected.
- **SameClass**: Indicates that two discourse units identified as CLASS are the same.

C Annotation Interface Details

Our annotation tool is we designed a simple and modularized annotation framework in 600 lines of JQuery, Javascript and HTML, with a Datastore backend[30]. Our annotation framework supports span annotation and relation tagging.

The annotation interface itself, shown in Figure 2, is powered by a stateful page object, called `PageHandler`, that is instantiated with several parameters (`page_height`, `buttons`, `relations`)

[30]Google Datastore is a NoSQL, scalable JSON store, which is suitable for our usecase. https://cloud.google.com/datastore

and handles all of the page interactions. The `PageHandler` is placed directly in the HTML page containing the text to be annotated, so any service that can render text can automatically become an annotation service. In our case, we built Jinja templates to render our HTML, since our server is coded in Python-Flask. We additionally provide a helper function that, with input data, can compile our Jinja templates as static, fully-functional AMT HTMLQuestions.

We use a Datastore backend to track progress towards annotation tasks, as shown in Figure **??**. We code data entries (the equivalent of MySQL tables) to track helper-statistics, `helper_summ`, how many tasks are left to assign, `incomp_tasks`, and how many annotations have been completed, `comp_annot`. We track these statistics to ensure that we can obtain multiple annotations for each task, and that no helper sees the same task more than once. We perform one GET request at the beginning of each user session to collect user-stats and then use client-side cookies throughout the session to minimize the number of requests we send to the back-end. We use a NoSQL database because they are low-latency and designed for streaming, and Datastore because our web-app is hosted on Google App Engine. We include our Datastore management back-end as part of the annotation package. To use our tool with other NoSQL providers,[31] a port is necessary.

D Website Design

In **Flow 1**, users can use a query box to perform full-text and faceted search on laws and then click on and return results to read the full text of the law. ElasticSearch powers both of these endpoints. This flow is useful for when journalists want to explore a specific term or concept irrespective of its discourse role, or simply familiarize themselves with the corpus.

In **Flow 2**, users can view aggregate counts of different discourse elements, by type, across the corpora. This helps to summarize the corpora from a functional standpoint, as described in Section 2. Users navigate this flow by clicking on one of five buttons to see the counts of each of the five principle discourse spans, then clicking on any of the returned span results to view all laws with this span. MySQL serves both of these endpoints (and provides additional metrics, such as a map in the `about.html` page, not shown here.).

In both flows, visitors can access our annotation framework, described in Section 3.2. From **Flow 1**, they can click search results to tag a specific paragraph, and from **Flow 2** they can click to correct an annotated paragraph. Additionally, they can annotate a randomly selected paragraph by clicking “Help Us Tag.”

[31] e.g. Amazon DynamoDB – `https://aws.amazon.com/dynamodb/`

E Additional Experimental Results

We give more results for the Span-Level tagging task, reporting on precision and recall as well as F1.

F Prompt Designs for GPT3.5

Here we give sample prompts, along with their true-label completions for each span.

F.1 Span Level Tagging Prompts

F.1.1 SUBJECT Identification

You are a legal assistant. I will show you a paragraph of law. Which entities gains powers, restrictions or responsibilities under this law? NOT which entities are used to test the law, or which entities are affected. In other words, what entity is the SUBJECT of the law? It might not aren’t always explicit, and sometimes can be expressed passively. Restrict your choices to an entity mentioned in the law OR "passive voice entity", if the entity is not explicitly mentioned in the text. Enumerate all instances of the entity in the text, even if repeated. If there is no entity that matches this description in the text, including "passive voice entity", say "no entity". If there are multiple segments of text in the law that apply, join them with a semi-colon. The order of text spans does NOT matter. Do NOT say anything else." I will give you 1 examples, and then you will perform the task yourself.

EXAMPLE: Law: "* 71. Special population census. The expenses incurred by a county, city, town, or village to conduct a special population census supervised by the United States bureau of the census pursuant to a contract made pursuant to section twenty of the general municipal law, three years." Answer: "no entity"

NOW IT’S YOUR TURN:

Law: "If a vacancy as described in subdivision (d)(1) occurs after the sixth Thursday before the primary election in any county having a metropolitan form of government with a population of more than five hundred thousand (500,000), according to the 2010 federal census or any subsequent federal census, then the members of the county executive committees who represent the precincts composing such senate district may nominate a candidate to appear on the November election ballot by any method authorized under the rules of the party." Answer:

»> vacancy as described in subdivision (d)(1); members of the county executive committees

F.1.2 EXCEPTION Identification

You are a legal assistant. I will show you a paragraph of law. What are exception cases when this law does not apply? Restrict your answer to text in the law. Join non-contiguous segments of text with a semi-colon. If there are no exception cases where this law does not apply, say "none". If there are multiple segments of text in the law that apply, join them with a semi-colon. The order of text spans does NOT matter. Do NOT say

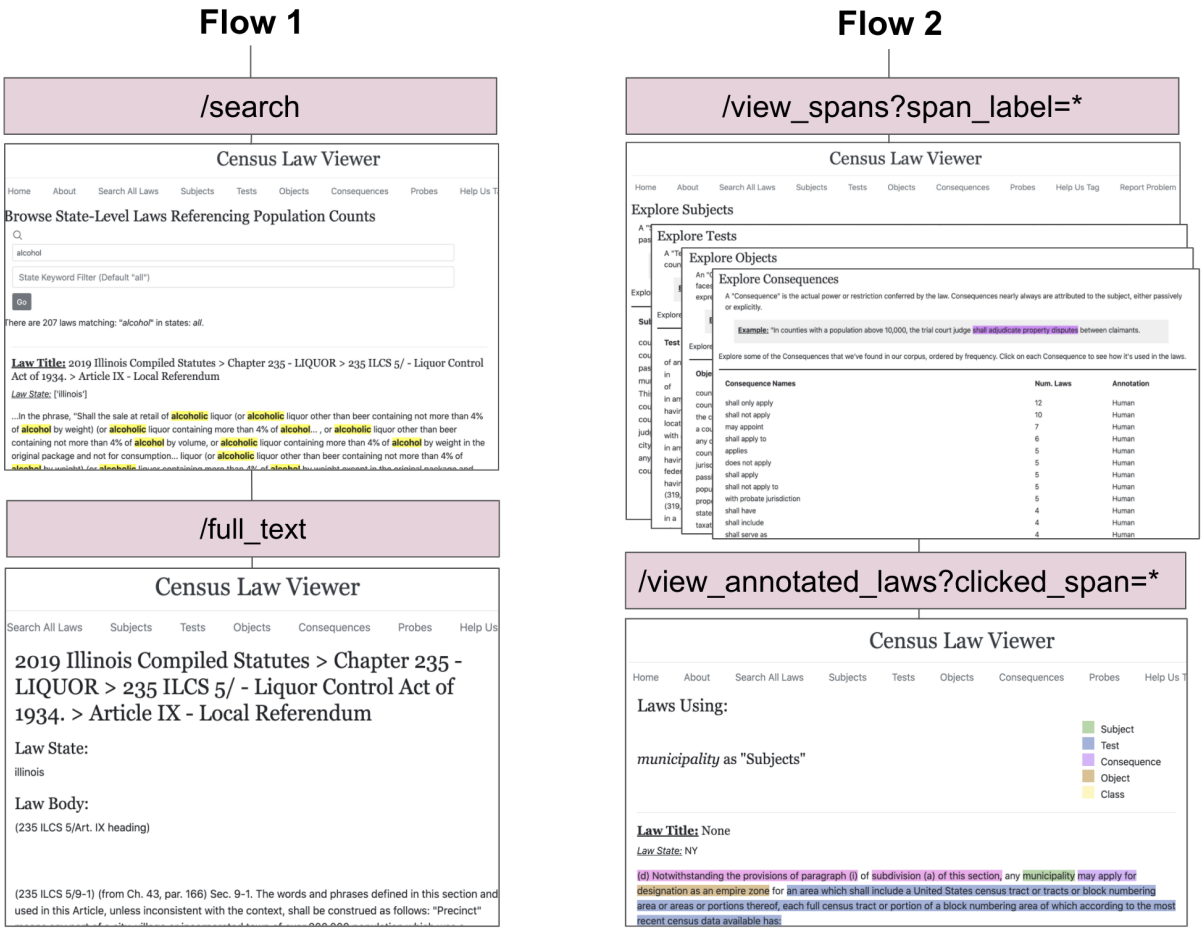

Figure 6: A flow-based sitemap for our website, `statecensuslaws.org`, with some details about the back-end and database setup. The left-column shows **Flow 1**, where a user can search and view full-text results. The right-column shows **Flow 2**, where a user can view top law-discourse spans and see all laws these spans are used in. Each flow leads to the annotation framework.

anything else." I will give you 1 examples, and then you will perform the task yourself.

EXAMPLE: Law: "Notwithstanding subdivision (b)(1), in counties having a population of not less than seventeen thousand two hundred fifty (17,250) nor more than seventeen thousand five hundred fifty (17,550), according to the 1990 federal census or any subsequent federal census, the budget committee shall be composed of six (6) members." Answer: "none"

NOW IT'S YOUR TURN:

Law: "In counties having a population of not less than three hundred nineteen thousand six hundred twenty-five (319,625) nor more than three hundred nineteen thousand seven hundred twenty-five (319,725), according to the 1980 federal census or any subsequent federal census, a library board of not less than seven (7) members nor more than nine (9) members may be appointed by the county legislative body and city governing bodies which are parties to the agreement, the number appointed by each to be determined according to the ratio of population in each participating city and in the county outside the city or cities, based on the most recent federal census; provided, that each shall appoint at least one (1) member." Answer:

»> provided, that each shall appoint at least one (1) member

F.1.3 TEST Identification

You are a legal assistant. I will show you a paragraph of law. Under what conditions does this law apply? In other words, what test is implied by the law? Restrict your answer to text in the law. Join non-contiguous segments of text with a semi-colon. If there are no conditions for this law to apply explicitly stated in the text, say "none". If there are multiple segments of text in the law that apply, join them with a semi-colon. The order of text spans does NOT matter. Do NOT say anything else." I will give you 1 examples, and then you will perform the task yourself.

EXAMPLE: Law: "(iii) Notwithstanding the foregoing, local governments and voluntary agencies shall be granted state aid of one hundred percent of the net operating costs expended by such localities and by voluntary agencies pursuant to contracts with such local governments or with the office of alcoholism and substance

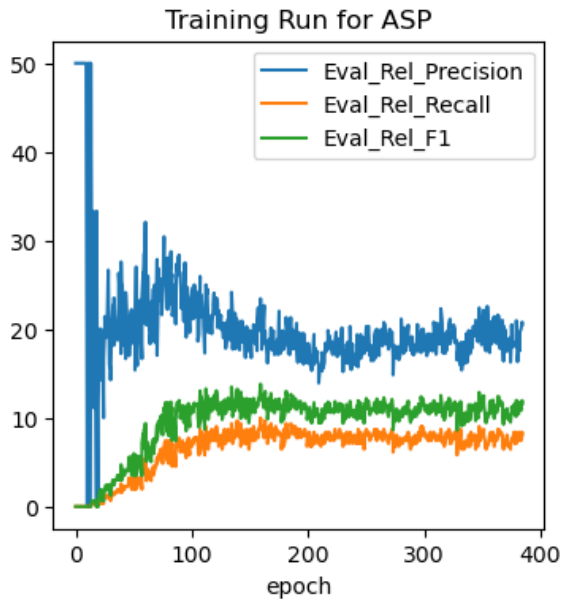

Figure 7: Training run for ASP. The ASP model is learning across epochs, it likely does not have enough data to fully distinguish the embedding space for jointly modeled task.

abuse services for alcohol crisis centers, chemical dependency programs for youth, residential services for recovering alcoholics and substance abusers and for alcoholism AIDS coordinators. Such state aid may also be granted to programs transferred from the task force on integrated projects for youth and chemical dependency. Such state aid shall also be granted for non-residential services determined to be necessary to serve the public interest by the commissioner of alcoholism and substance abuse services provided by local governments having a population of one hundred twenty-five thousand or less as determined by the last preceding federal census, or by voluntary agencies pursuant to contracts with such local governments." Answer: "determined to be necessary to serve the public interest by the commissioner of alcoholism and substance abuse services; provided by; having a population of one hundred twenty-five thousand or less as determined by the last preceding federal census; pursuant to contracts with; with; transferred from the task force on integrated projects for youth and chemical dependency"

NOW IT'S YOUR TURN:

Law: "(2) If two or more counties included in the measure are required to prepare a translation of ballot materials into the same language other than English, the county that contains the largest population, as determined by the most recent federal decennial census, among those counties that are required to prepare a translation of ballot materials into the same language other than English shall prepare the translation, or authorize the authority to prepare the translation, and that translation shall be used by the other county or counties, as applicable." Answer:

»> are required to prepare a translation of ballot materials into the same language other than English; that contains the largest population, as determined by the most recent federal decennial census; among those counties that are required to prepare a translation of ballot materials into the same language other than English

F.1.4 OBJECT Identification

You are a legal assistant. I will show you a paragraph of law. Which entities are affected by the powers of this law? NOT which entities gain powers, but who is affected by those in power? In other words, what entity is the object of the law? It might not aren't always explicit, and sometimes can be expressed passively. Restrict your choices to an entity mentioned in the law OR "passive voice entity", if the entity is not explicitly mentioned in the text. Enumerate all instances of the entity in the text, even if repeated. If there is no entity that matches this description in the text, including "passive voice entity", say "no entity". If there are multiple segments of text in the law that apply, join them with a semi-colon. The order of text spans does NOT matter. Do NOT say anything else." I will give you 1 examples, and then you will perform the task yourself.

EXAMPLE: Law: "This subsection (b) applies only to counties with a metropolitan form of government and to counties having the following populations according to the 1970 federal census or any subsequent federal census:" Answer: "no entity"

NOW IT'S YOUR TURN:

Law: "An authority shall not initiate any redevelopment project under this chapter until the governing body, or agency designated by it or empowered by law so to act, of each city or town, herein called "municipalities," and any county having a population of not less than two hundred seventy-five thousand (275,000) nor more than three hundred twenty-five thousand (325,000), according to the 1980 federal census or any subsequent federal census, in which any of the area to be covered by the project is situated, has approved a plan, herein called the "redevelopment plan", which provides an outline for the development or redevelopment of the area and is sufficiently complete, to:" Answer:

»> any redevelopment project under this chapter

F.1.5 PROBE Identification

You are a legal assistant. I will show you a paragraph of law. Which entities are used to determine when this law applies? NOT which entities gain powers, OR is affected by the law. In other words, which entities are probed by the law? Restrict your answer to text in the law. Join non-contiguous segments of text with a semi-colon. If there is no entity that matches this description in the text, including "passive voice entity", say "no entity". If there are multiple segments of text in the law that apply, join them with a semi-colon. The order of text spans does NOT matter. Do NOT say anything

else." I will give you 1 examples, and then you will perform the task yourself.

EXAMPLE: Law: "2. The commissioner is authorized to contract to make a state grant, within the limit of appropriation therefor, to any planning unit for up to ninety percent of the costs to prepare, update or revise its local solid waste management plan; provided, however, that no such grant has been previously made to a planning unit which is a part of or is served by the planning unit seeking such grant. A planning unit may receive a grant pursuant to this subdivision which shall not exceed the greater of twenty-five thousand dollars or one dollar for each resident of the planning unit, based upon the current federal decennial census." Answer: "no such grant; a planning unit; the planning unit"

NOW IT'S YOUR TURN:

Law: "Such functions may also be delegated by the municipality to any not-for-profit corporation acting for or on behalf of such municipality; provided, that, except in any county with a metropolitan form of government and having a population of four hundred thousand (400,000) or more, according to the 1980 federal census or any subsequent federal census, the site selection for an energy production facility may be delegated to any such not-for-profit corporation, but shall be subject to the approval by a two-thirds (2/3) vote of the legislative bodies of the city and the county in which such city is located for whom or on whose behalf such not-for-profit corporation is acting prior to the purchase of any such site." Answer:

»> any county

F.1.6 CONSEQUENCE Identification

You are a legal assistant. I will show you a paragraph of law. What are the powers or obligations granted by this law? In other words, what is the law's consequence? Restrict your answer to text in the law. Join non-contiguous segments of text with a semi-colon. If there are no powers or obligations explicity stated in the text, say "none". If there are multiple segments of text in the law that apply, join them with a semi-colon. The order of text spans does NOT matter. Do NOT say anything else." I will give you 1 examples, and then you will perform the task yourself.

EXAMPLE: Law: "After January 1, 1980, with respect to the construction, purchase, or lease of buildings which are located or will be located in a standard metropolitan statistical area (SMSA) with a population of 250,000 or more according to the most recent decennial census, which is served by a public transit operator, as defined in Section 99210 of the Public Utilities Code, the board shall give consideration to the location of existing public transit corridors, as defined in Section 50093.5 of the Health and Safety Code, for the area. Construction, purchase, or lease of buildings at locations outside of existing public transit corridors may be approved after the board has determined: (1) the purpose of the facility does not require transit access; or (2) it is not feasible to locate the facility in an existing transit corridor; or (3) the transit operator will provide service as needed to effectively serve the facility. The board may request the assistance of the transit operator in making its determination and shall notify the operator of its decision." Answer: "may be approved; shall give consideration to; may request the assistance of; in making its determination; shall notify; of its decision"

NOW IT'S YOUR TURN:

Law: "This part only applies in those counties with a metropolitan form of government and in those counties with a population according to the 1970 federal census or any subsequent federal census of:" Answer:

»> applies

Discourse Unit	Example
SUBJECT	clerk and master legislative body any person the board of mayor and aldermen “Club”
OBJECT	the library and recreational facilities. to the electors of the county presiding officer the property owners the tenants and their property and the safety and the protection of the premises.
TEST	having a population of not less than eight hundred twenty-five thousand (825,000) nor more than eight hundred thirty thousand (830,000)... upon adoption of a resolution by a two-thirds (2/3) vote of the county legislative body authorizing the county trustee to collect delinquent property taxes as provided in this subsection who owns real property situated within the corporate limits of such municipality upon entering an order finding it in the best interest of judicial efficiency in areas of historical significance to a locality, the county and the state
CONS.	shall, upon collection of state fines and costs, remit such fines and costs to may be levied be governed by shall make eligible for the waiver be paid from the same fund used for maintaining and operating the county free library.
EXCEPTION	wherever its disapproval of a redevelopment project has been dissolved as prescribed by contracting with other counties and/or cities for joint operation of a free public library except the clerk of the supreme court and chief deputy clerks of the supreme court provided, that each shall appoint at least one (1) member unless the board of supervisors of the county shall, by resolution, provide for fees in excess of that amount
PROBE	county enrolled member and spouse city in Canada an enrolled member of an incorporated volunteer fire company, fire department or incorporated voluntary ambulance service private acts of the state
CLASS	the superior [court] for such county general sessions court [clerk] [the legislative body] of the municipality. the mental health [court] [the commissioner] of mental health,
DEFINITION	shall be determined by the last federal decennial or local special population census... is the same proportion of the total population of the district as each of the other areas. that is the sum of the county public hospital health system’s gross inpatient revenue shall include The Municipality of Metropolitan Toronto and any other similar corporation in Canada means any regular and full-time employee of a county with a metropolitan government

Table 6: Example spans from each discourse type in our annotated dataset.

Source Span	Target Span	Permissible Relations
OBJECT	CLASS	hasProperty
	OBJECT	continuation, And, Or, sameEntity, By, To
	CONSEQUENCE	entityEmpoweredTo, entityRequiredTo
	TEST	entityTested
	DEFINITION	definedAs
	SUBJECT	sameEntity, And, Or, Of
SUBJECT	CLASS	hasProperty
	OBJECT	continuation, And, Or, sameEntity
	CONSEQUENCE	entityEmpoweredTo, entityRequiredTo
	TEST	entityTested
	DEFINITION	isDefinedAs
	SUBJECT	sameEntity, And, Or, Of
TEST	CONSEQUENCE	conditionalConsequence
	TEST	continuation, And, Or, followedBy
	SUBJECT	testConcerns
	PROBE	testConcerns
	EXCEPTION	exceptedBy
CONSEQUENCE	CLASS	hasProperty
	OBJECT	Affects, comparison
	CONSEQUENCE	continuation, And, Or, followedBy
	TEST	conditionedBy
	DEFINITION	Affects
	SUBJECT	Affects, comparison
	EXCEPTION	conditionedBy
CLASS	CLASS	continuation, And, Or, sameClass
	OBJECT	propertyOf
	DEFINITION	definedAs
	SUBJECT	propertyOf
	PROBE	propertyOf
EXCEPTION	OBJECT	excepts
	CONSEQUENCE	excepts
	TEST	excepts
	SUBJECT	excepts
	PROBE	excepts
	EXCEPTION	excepts, continuation, And, Or
PROBE	CLASS	hasProperty
	TEST	entityTested
	PROBE	sameEntity, And, Or, Of
DEFINITION	CLASS	defines
	OBJECT	defines
	DEFINITION	continuation
	SUBJECT	defines
	PROBE	defines

Table 7: All possible relations between discourse units identified in our span-tagging process.

	SUBJECT			CONSEQUENCE			OBJECT		
	P	R	F1	P	R	F1	P	R	F1
0-shot	39.5	30.5	34.4	12.5	7.9	9.7	13.9	15.7	14.8
3-shot	32.1	31.3	31.7	22.4	24.3	23.3	18.3	23.1	20.4
5-shot	27.9	34.1	30.7	23.2	25.1	24.1	13.9	18.4	15.9
8-shot	27.4	32.5	29.7	21.6	25.5	23.4	13.4	19.2	15.8
fine-tuned	41.2	43.1	42.1	51.0	48.8	49.9	38.8	33.3	35.9

Table 8: Precision, Recall and F1 for the first three discourse tags we studied.

	PROBE			TEST			EXCEPTION		
	P	R	F1	P	R	F1	P	R	F1
0-shot	11.3	16.7	13.4	42.8	30.2	35.4	53.3	56.1	54.7
3-shot	23.2	36.0	28.2	45.0	42.8	43.9	45.0	47.4	46.2
5-shot	26.7	36.4	30.8	48.8	50.9	49.8	41.8	49.1	45.2
8-shot	29.1	39.6	33.5	46.7	50.2	48.4	51.6	56.1	53.8
fine-tuned	38.8	31.7	34.9	55.2	51.1	53.0	51.5	61.4	56.0

Table 9: Precision, Recall and F1 for the last three discourse tags we studied.

	Macro			Micro		
	P	R	F1	P	R	F1
0-shot	28.9	26.2	27.1	25.0	21.4	22.7
3-shot	31.0	34.2	32.3	28.8	32.1	30.1
5-shot	30.4	35.7	32.8	28.6	33.6	30.8
8-shot	31.6	37.2	34.1	28.5	34.2	31.0
fine-tuned	46.1	44.9	45.3	45.6	43.3	44.3

Table 10: Macro-average and Micro-averaged Precision, Recall and F1 for all discourse tags we studied.