Definition Disambiguation In The U.S. State Regulations

Dylan Walker, Chuntian Chi, & Haonan Peng

Agenda

Legal Information Institute

- Introduction
  - US State Regulations
  - Previous Work and Background
  - What is Scoping?
- Problem Overview
  - Task Breakdown
- Definitions
  - Definition Insights
  - Terms with Multiple Definitions
- Scope Identification and Disambiguation
  - Identifying Scoping Language
  - Handling Disambiguation

Agenda (Continued)

Legal Information Institute

- XML Markup
  - Data Representation
  - Assigning Definitions
- Future Work
- Conclusions

Introduction

US State Regulations

Legal Information Institute

- The US State Regulations are the official legal regulations for each of the 50 states in the US.
- Each individual regulation is developed by the legislative body of its respective state and acts as the official codification of laws that govern that state

Cornell Legal Information Institute

- The Legal Information Institute (LII) at Cornell University is an organization that publishes many legal corpora including the US State Regulations.
- LII works to increase accessibility to legal information by developing systems to empower users from outside the legal profession to more easily access and understand the laws that govern them.

Why is Definition Disambiguation Important?

- A term is defined with different definitions across and within states.

Legal Information Institute

State	Term	Definition
Florida	Equine	Any member of the family Equidae, including horses, mules, asses, and zebras.
North Carolina	Equine	Any member of the equine family, including horses, ponies, mules, asses and other equines;
Kansas (title 25)	Written notice	a written notification which is either hand-delivered, facsimile-transmitted or sent by certified mail.
Kansas (title 23)	Written notice	any paper or electronic document relevant to a matter that is filed with the Commission.

Previous Works

Legal Information Institute

- Data
  - XML format
  - State regulations for all 50 states
- Definition Extraction
  - CSV format
  - Where terms are defined

idnum	datapath	Tag	Method	Term	Definition
Agency 100 - Article 25 - section 1	filepath	subsect	cues	Office	any place intended for the practice of the healing arts in the state of Kansas.

- We want to know where they apply

What is Scoping?

Legal Information Institute

- The **scope** of a definition denotes the segments of the corpus for which that definition is to be applied
- Every defined term has an associated scope that designates that it's usage within the regulation
- These scopes vary in range from single parts or sections to entire titles

Title 1 - Division 1 - Chapter 1 - Article 1 - Section Cal-Code-Regs-Tit-1

The following definitions shall apply to the regulations contained in this chapter:

(1) "APA" means the part of the California Administrative Procedure Act appearing in California Government Code, Title 2, division 3, part 1, chapter 3.5, commencing with section 11340, which generally governs the adoption, amendment, or repeal of regulations by California state agencies.

Goals

Legal Information Institute

- The end goal is to provide users with correct in-text annotations of term definitions within their context.
- Our team's goals are:
  - Develop methods to disambiguate definitions of a term
  - Associate a term with its correct definition in its context
  - Annotate term definitions in regulations

Problem Overview

Previous Work - Definition Extraction

Legal Information Institute

- Previous group of students developed a rule-based system to extract definitions from the US State Regulations
- Same term defined with multiple definitions in different scopes
- Different terms defined with similar definitions in different scopes

Previous work Usage

Legal Information Institute

- Well-defined scoping language

10666	0	title:45a:ur/regulation	subsect	cues	conservatc or her affairs," incapable of caring for oneself," "respondent," and "ward" are defined in accordance with C.G.S. Section 45a-644 .

	Term	Filepath	IDNum	Scoping Phrase	Scope
763	conservatc	States/con	title:45a:unprefixed:651::	as defined in this section	section:Conn-Agencies-Regs-SS-45a-651-1

- Used to generate insights of different corpora
- Locate the context around the extracted definition


```
{'title:01:subtitle:01:chapter:01.01:section:Md-Code-Regs-01-01-1972-03': {'Act': [" of Article 66-1/2 of the Annotated Code of Maryland (1970 Replacement Volume), and for purposes of the Federal Highway Safety Acts of 1966 and 1970, the Secretary of Transportation, is hereby designated as the Governor's Representative for Highway Safety."],
  'Department': [" There has been established within the Department of Transportation of the Executive Branch of the State Government a Division of Transportation Safety within which the Office of Highway Safety Coordinator shall be incorporated. The powers and authority hereby delegated to the Secretary of Transportation as the Governor's Representative, may be redelegated by him to the Division of Transportation Safety at his discretion."],
```

Task Breakdown

Legal Information Institute

Gathering Statistics

- Collect statistics on term frequency and diversity
- Measure the frequency and diversity of scoping language
- Compare and contrast statistics across state corpora

Scoping

- Establish a baseline set of scoping phrases
- Derive a system for identifying scoping language in XML
- Disambiguate scoping language

XML Markup

- Identify defined term instances in XML
- Create a unique “definiendum” XML tag for definition markup
- Create a unique identifier to reference each definition
- Assign definitions to term instances using scope
- Markup term instances in XML

Definitions

Definitions Overview

Legal Information Institute

- Definitions within legal corpora can be tricky and often differ from colloquial meanings

State	Classification of zebras
South Carolina	For the purpose of these regulations, "horse" means any member of the equine family including horses, mules, asses, zebras or other equadae.
Maryland	"Horse" means any member of the equine family, including horses, mules, asses, zebra, or other equidae.
Utah	Equine - means any animal in the family Equidae, including horses, asses, mules, ponies, and Zebras

Definition Insights

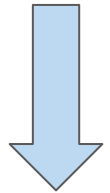
- On average for each state corpus 31.99% of all terms are defined more than once
  - “Commissioner” defined 314 times in Connecticut - most of any defined term
  - “Department” is the most defined term in 21 states
    - 154 times in New Mexico
    - 132 times in Alaska
    - 9 times in Missouri
- Only 38.94% of all terms are accompanied by scoping language

Terms with Multiple Definitions

Legal Information Institute

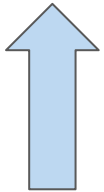
Part-of-Speech (POS) Tag Exploration

- Assume the POS tag of ambiguous words can clearly determine the correct sense
- Facilitate the extraction of truly legally-defined occurrences of terms



§ 1.1 General.

(a) The provisions of regulations promulgated under the Federal Food, Drug, and Cosmetic Act with respect to the doing of any act shall be applicable also to the causing of such act to be done.



Terms with Multiple Definitions

Legal Information Institute

Part-of-Speech (POS) Tag Exploration

- Focused on two main POS tags for analysis
- Three types of tags in the program:
  - Noun
  - Verb
  - Unk (Unknown)

POS tags for Spacy Library

POS	DESCRIPTION	EXAMPLES
ADJ	adjective	*big, old, green, incomprehensible, first*
ADP	adposition	*in, to, during*
ADV	adverb	*very, tomorrow, down, where, there*
AUX	auxiliary	*is, has (done), will (do), should (do)*
CONJ	conjunction	*and, or, but*
CCONJ	coordinating conjunction	*and, or, but*
DET	determiner	*a, an, the*
INTJ	interjection	*psst, ouch, bravo, hello*
NOUN	noun	*girl, cat, tree, air, beauty*
NUM	numeral	*1, 2017, one, seventy-seven, IV, MMXIV*
PART	particle	*’s, not,*
PRON	pronoun	*I, you, he, she, myself, themselves, somebody*
PROPN	proper noun	*Mary, John, London, NATO, HBO*
PUNCT	punctuation	*., (,), ?*
SCONJ	subordinating conjunction	*if, while, that*
SYM	symbol	*$, %, §, ©, +, −, ×, ÷, =, :), 😝*
VERB	verb	*run, runs, running, eat, ate, eating*
X	other	*sfpksdpsxmsa*
SPACE	space	

Terms with Multiple Definitions

Legal Information Institute

Word Vector Similarity Comparison

- In-scope vs. Out-scope

Florida Corpus ***Chapter 68B-12***

```
▼<codetext>
   <para>For the purpose of this chapter, except where the context clearly requires otherwise:</para>
▼<subsect>
   <designator>(5)</designator>
   "King mackerel," also commonly referred to as "kingfish," means any fish of the species Scomberomorus cavalla.
 </subsect>
```

Terms with Multiple Definitions

Legal Information Institute

Word Vector Similarity Comparison

- In-scope vs. Out-scope

```
▼<subsect>
   <designator>(5)</designator>
   "King mackerel," also commonly referred to as "kingfish," means any fish of the species Scomberomorus cavalla.
 </subsect>
```

Extracted Definition of “kingfish”: Any fish of the species Scomberomorus cavalla.

Context of “kingfish” in ***Chapter 68B-12*** : “King mackerel,” also commonly referred to as “kingfish,” means any fish of the species Scomberomorus cavalla.

Terms with Multiple Definitions

Legal Information Institute

Word Vector Similarity Comparison

- In-scope vs. Out-scope

Context of “kingfish” in ***Chapter 68B-12*** : “King mackerel,” also commonly referred to as “kingfish,” means any fish of the species Scomberomorus cavalla.

Program Disambiguated Sense of “kingfish”

Terms with Multiple Definitions

Legal Information Institute

Word Vector Similarity Comparison

- In-scope vs. Out-scope

Similarity Comparison using Word Vector

Extracted Definition

Program Disambiguated Sense of term “kingfish”

Part-of-Speech (POS) Tag Exploration

Legal Information Institute

- Use Part-of-Speech tag of the defined terms to find the right scope

```
# for txt in all_texts:
corpus_len = corpus.shape[0]
i = 0

while i < corpus_len and total_term_length <=20:
# for i in range(corpus_len):
  txt = corpus['Text'][i]
  i+=1

  for word in txt.split(" "):

    if word in unique_term_lst:

      for token in nlp(txt):

        if str(token) == word:
          total_term_length +=1
          print(token.pos_)
          if token.pos_ in nouns:
            pos_count['N'] +=1
          elif token.pos_ in verbs:
            pos_count['V'] +=1
          else:
            pos_count['unk'] +=1
```

```
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
PROPN
NOUN
NOUN
NOUN
NOUN
NOUN
NOUN
```

Word Vector Similarity Comparison

Legal Information Institute

- Use disambiguated context of a defined term to compare with extracted definitions
- For example, if term A is defined in article 15, a disambiguated definition will first be generated for that term using its context. Then, the proper definition of that term will be extracted from article 15. To contrast with the In-Scope definition of term A, another definition of term A will be extracted from an article that is not article 15.
- Total Noun POS (Part-of-Speech) occurrences account for 52.5% of all defined term occurrences.
- Total Verb POS (Part-of-Speech) occurrences account for only 2.8% of all defined term occurrences.

Challenge - To define or not to define

Legal Information Institute

- POS tags are mostly Noun tag or Unk (unknown) tag
  - Total Noun tag occurrences account for 52.5% of all defined term occurrences.
  - Total Verb tag occurrences account for only 2.8% of all defined term occurrences.
  - Total Unk (unknown) tag occurrences account for 44.7% of all defined term occurrences.
- In-scope and Out-scope similarity are close
  - Average In-scope Similarity is about 73.9%
  - Average Out-scope Similarity is about 73.5%

Scope Identification and Disambiguation

Predefined Scoping Phrases

Legal Information Institute

- Wide variety of corpus hierarchy and scoping language across states
- Calls for a predefined list of generalizable scoping phrases
- Collected a list of ~20 template phrases through exploration

Example Scoping Phrases

When used in
As used in this
Under the provisions of this
For purposes of this
Within the scope of this
For purposes of the application under this
As defined in
As specified in

Identifying Scoping Language

Legal Information Institute

- Extract definitions from each State Regulation
- Use extracted definitions to identify all files containing definitions
- Search each definition file for predefined scoping phrases to identify scoping sentences
- Search scoping sentences for scope keywords (Ex. “Chapter”, “Title”, etc)

Ambiguous Scopes

Legal Information Institute

- A piece of scoping language is **well-defined** if the range of the scope is clearly stated and easily interpretable.

> *"As used in this section: 'Commissioner' means the Commissioner of Motor Vehicles or his authorized representative."* - ***Connecticut, Title 1 Section 1H-8***

- A piece of scoping language is **ambiguous** if the range of the scope is not easily interpretable and requires additional context and information than the scoping language itself.

> *"As used in these rules: 'Department' means the Michigan department of health and human services."* - ***Michigan, Licensing and Regulatory Affairs***

Handling Scopes

Legal Information Institute

- Well defined scoping language can be easily resolved by keyword detection
- Ambiguous scoping language must be resolved to a default scope
- Definitions absent of any scoping language also assigned a default scope

Location	Scoping Phrase	Scope
Title 1 - Chapter 7 - Section 3	“For purposes of this chapter”	*Title 1 - Chapter 7*
Title 2 - Division 1 - Article 4	“As used in these rules”	*Title 2 - Division 1 - Article 4*
Title 3 - Part 1 - Subpart C	No scoping language detected	*Title 3 - Part 1 - Subpart C*

XML Markup

Markup Overview

Goals

- Find all defined term instances in a given state regulation
- Modify XML documents to indicate defined terms
- Include unambiguous references to corresponding definitions

Standards

- All Markup definitions should use the “definiendum” tag
- XML structure should be maintained for all non-definition elements
- Spacing and content should be consistent between original and markup documents

The Definiendum Tag

Legal Information Institute

During the markup process we need a unique tag that denotes our defined terms and any necessary information for linking them with their associated definitions. To do this we create the special “definiendum” tag. This tag includes three main attributes:

1. ***ID*** - unique identifier of the corresponding definition
2. ***numOccur*** - Which number occurrence of the definition the current instance is
3. ***Markup = “no”*** - Denotes a definition was matched using a default scope.

<designator>(e) </designator> “<definiendum id=“34981479388646131 23” numOccur=“1” markup=“no”>Residential Program</definiendum>” means a duly licensed, certified or approved foster family boarding home, agency boarding home, supervised independent living program, group home, group residence, or any combination thereof as such terms are defined under 18 NYCRR section.

Representing Defined Terms

Legal Information Institute

- A **defined term** describe a unique definition within a given state corpus.
- Let's look at an example defined term from Florida

Attribute	Value
Term	Pompano
Definition	"any fish of the species Trachinotus carolinus, or any part thereof."
Definition ID	8008162398517937498
Scope	*Department 68 - Division 68B - Chapter 68B-35*
Number of Occurrences	7

Representing Term Instances

Legal Information Institute

- A **term instance** is a unique appearance within the text of a given defined term.
- These are the exact elements we are looking to markup! Here's an example from the Ohio corpus

Attribute	Value
Term	honey
Definition	"the nectar and saccharine exudation of plants that has been gathered, modified, and stored in a honeycomb by honeybees."
Definition ID	36791096840139527 24
Instance Number	3

Identifying Term Instances

Legal Information Institute

- Apply definition extraction and scope assignment to all defined terms
- Combine all defined terms into a dictionary structure
  - Keys correspond to unique term strings
  - Values are lists of DefinedTerm objects for each possible definition of the term
- Cross-compare XML document text with defined terms using modified string matching

Example Dictionary Key	Example Dictionary Value
Applicant	1. **Definition:** A person seeking certification as an officer. **Scope**: *Agency 106 - Article 2* 2. **Definition**: A landowner or legal agent applying for financial assistance to construct or apply conservation or pollution control practices. **Scope**: *Agency 11 - Article 1*

Assigning Definitions

Legal Information Institute

1. Identify all terms that have a definition that is in-scope for this document

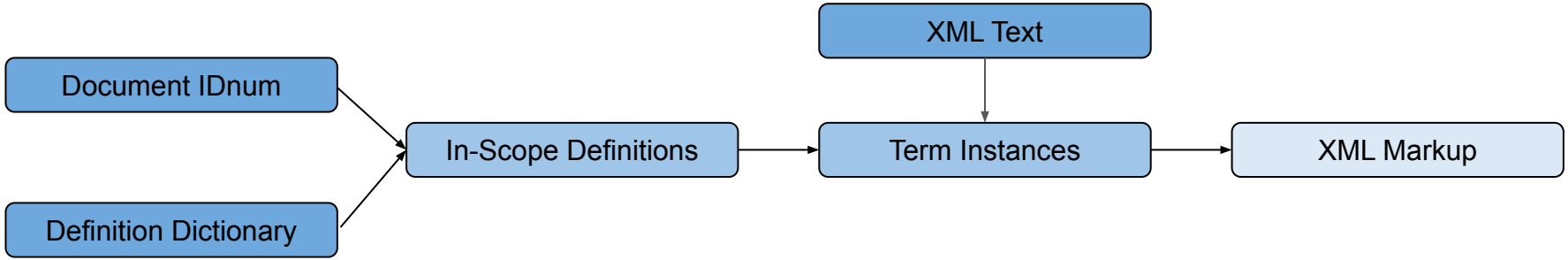
   Document: Agency 28 - Article 50

Example Dictionary Key	Example Dictionary Value
Applicant	1. **Definition:** A person seeking certification as an officer. **Scope**: *Agency 106 - Article 2* 2. **Definition**: A landowner or legal agent applying for financial assistance to construct or apply conservation or pollution control practices. **Scope**: *Agency 11 - Article 1*
Applicator	1. **Definition**: a structure that determines the extent of the treatment field at a given distance from the virtual source. **Scope**: Agency 28

Assigning Definitions

Legal Information Institute

1. Identify all terms that have a definition that is in-scope for this document
2. For each element of the document: search for matches between the element text and the in-scope definitions
3. For each match assign definition based on matching scope
4. Markup each match with “definiendum” tag



Markup Challenges

Challenges - Lemmatization

Legal Information Institute

Lemmatization: the grouping together of different forms of the same word.

- Lemmatized term
- List of tokenized words from text
- List of lemmatized tokens
- List of positions of words in original text (0-index based)

Example:

Term: facility

Tokens: ['fire', 'protection', 'and', 'public', 'safety', 'facilities', ';']

Lemmatized: ['fire', 'protection', 'and', 'public', 'safety', 'facility', ';']

Position: [[(1, 5)], [(10, 20)], [(21, 24)], [(25, 31)], [(32, 38)], [(39, 49)], [(49, 50)]]

Challenges - Subtoken Definitions

Legal Information Institute

- Naive string matching creates problems with subtoken definitions
- First approach was to add spacing to either side of a term during search
    - Fails to resolve punctuation and sentence/element boundaries
- Solved issue by checking if boundary contained alphanumeric characters

<designator> (A) </designator> During the 18-month period immediately p<definiendum id="3510363150165908518" numOccur="1">rec</definiendum>eding the license expiration date, the <definiendum id="03461673500039744478" numOccur="1">person</definiendum> completed at least 50 credits

Challenges - Nested Definitions

Legal Information Institute

- Nested instances occur when a defined term appears as a word of phrase within the use of another defined term
- Introduces complications to both identification/definition assignment as well as markup

Defined Term	Definition
Attendant	a staff person or volunteer who provides direct supervision of a juvenile.
Attendant Care	one-on-one direct supervision of a juvenile who has been taken into custody. Attendant care shall not exceed 24 hours exclusive of weekends and court holidays.
Attendant Care Facility	a boarding home for children at which attendant care is provided.

Challenges - Nested Definitions

Legal Information Institute

- Only the outermost defined term of a nested instance shall be marked up
- Sorted defined terms from longest to shortest to ensure outermost terms are processed first
- Prevent searching from within already tagged “definiendum” elements

"<definiendum id="34586237776468265 28" numOccur="1" markup="no">Attendant care</definiendum>" means one-on-one direct supervision of a juvenile who has been taken into custody.

Future Work

Future Work

Legal Information Institute

- More sophisticated scope extraction
  - Potentially use machine learning techniques to improve scoping
  - Handle ambiguous scopes - derive rules from domain expert
- Improve definition assignment to include more than scope
- Improve efficiency of definition extraction and assignment

Conclusions

Conclusions

Legal Information Institute

- Collected insightful statistics on defined terms in the US State Regulations
- Applied a variety of methods to identify and disambiguate scoping language
- Developed a system to assign definitions to term instances
    - Marked up all the defined terms in the xml file
        - Terms with a well-defined scoping language
        - Terms with an ambiguous scoping langage
        - Terms without any scoping language

Thank You!