Finding Contextually Consistent Information Units in Legal Text

Dominic Seyler (dseyler2@Illinois.edu)

ChengXiang Zhai (czhai@Illinois.edu)

University of Illinois at Urbana-Champaign

Paul Bruin (paul.bruin@regology.com)

Pavan Bayyapu (pavan.bayyapu@regology.com)

Regology

Motivation

- Terms in the laws of a legislature can be highly contextual.
- 26 USC § 7701(a)(1): for purposes of Title 26 “The term ‘person’ shall be construed to mean and include an individual, a trust, (...) or corporation.”
- 42 USC § 2000e(a): for purposes of the subchapter “The term ‘person’ includes one or more individuals, governments, (...) or receivers.”
- Context is not identical across the corpus.

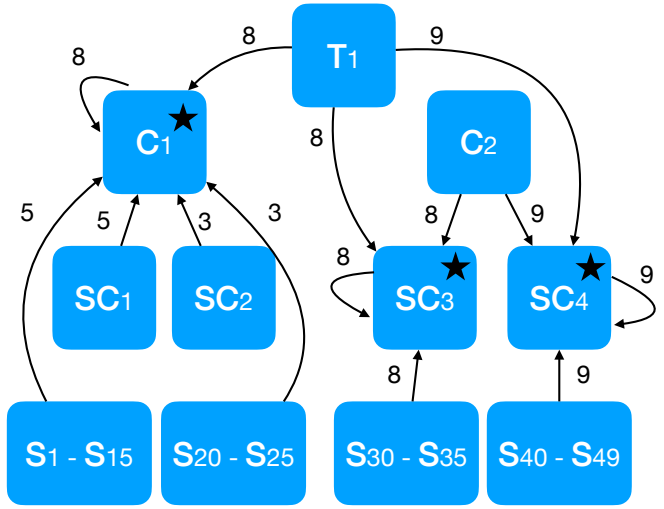
Goal

- Assist professionals when reading legal text by finding contextually consistent information units.
- Our method is modeled to emulate the “contextualization process”:

1. Reader notices that definition of ‘person’ is not in the current section.
2. Reader needs to find the definition of ‘person’ that is applicable to the current section.
3. Reader needs to understand which other sections the definition applies to.

What is a Root Context?

- We call contextually consistent information units Root Contexts.
- Commonly represents an individual law on a specific topic.
- Root context is used where a codified corpus' hierarchical structure does not designate a single level for individual laws.

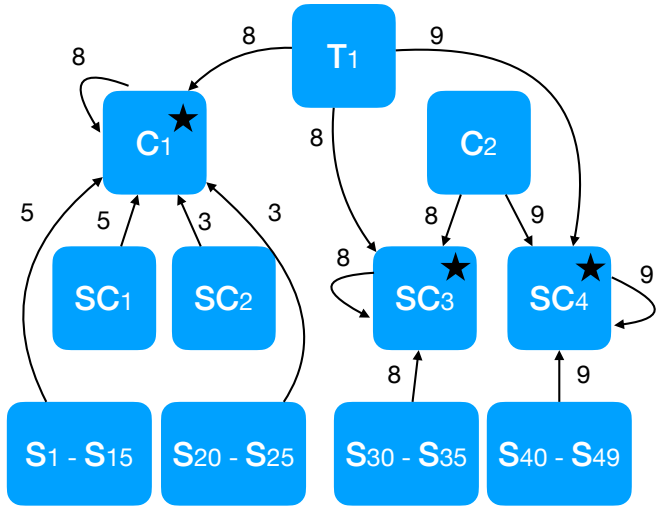


Approach


Build the hierarchy-reference graph
Find root context indicators
Traverse hierarchy-reference graph

Build the hierarchy-reference graph

- Nodes in the graph are “hierarchy branches”, e.g., (Title 1, Chapter 1).
- Find hierarchy references in section text using regex, e.g., “this (<hierarchy level>)“.
- Add weighted edges that symbolize hierarchy references.
- Aggregate counts for each hierarchy level by moving “upwards”.



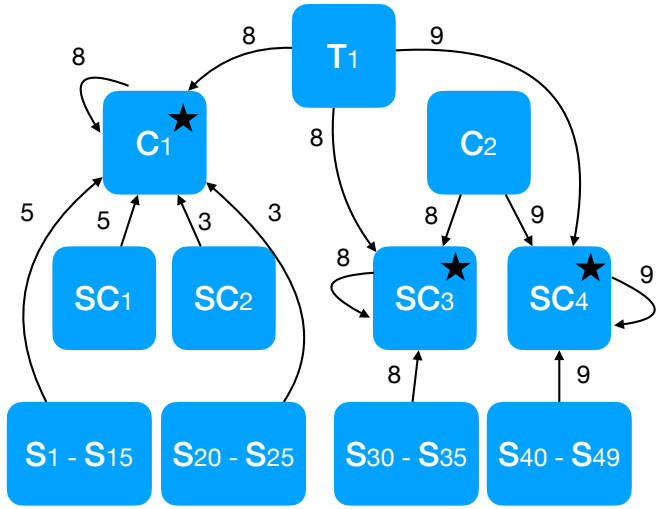
Find root context indicators

- References in certain sections that contain definitions or purposes carry more weight than regular sections.
- We define a set of regular expressions that extract the hierarchy references in these sections.
- If one of the regular expressions matches the text within a section, the hierarchy branch of the extracted reference is considered a root context.

Target	Regular Expression
definitions	this (\w+) (?:describes|sets forth|governs|contains)
definitions	the following definitions apply to this (\w+)
purposes	the purposes? of (?:the)?.*this (\w+) (?:are|include|is)
purposes	this (\w+) sets forth

Traverse hierarchy-reference graph

- For every node in the graph, we perform multiple hops following the outgoing edge with the highest weight.
- If a node points to itself, we consider the node a root context.
- If a node has no outgoing edges, we move to the node's parent, and continue the procedure.



Experimental Setup

- Corpora:

1. United States Code (USC)
2. California Law (CACL)
3. Texas Statutes (TXST)
4. Illinois Compiled Statutes (ILCS)
5. Consolidated Laws of New York (NYCL)

Dataset	Total	Number Root Contexts	%
USC	166,086	3,040	1.83
CACL	177,862	2,887	1.62
TXST	239,259	4,419	1.84
ILCS	19,088	800	4.19
NYCL	4,201	306	7.28

- Expert annotations: Label each node as a root context (“1” label) or not (“0” label).
- Metrics: $F1$-score, precision, recall of the class under consideration (i.e., “1” label) and classification accuracy

Experimental Results

- Generally achieves high precision (≥ 0.95)
- Recall is also high for state corpora
- Less Recall for our federal corpus
- Accuracy is close to 1 for all corpora

Dataset	F_1	Precision	Recall	Accuracy
USC	0.71	0.95	0.56	0.99
CACL	0.97	0.98	0.97	1.00
TXST	0.95	0.98	0.91	1.00
ILCS	0.98	0.99	0.97	0.99
NYCL	0.92	1.00	0.85	0.99

Conclusions and Impact

- Introduced problem of finding contextually consistent information units and developed methodology to find these units.
- Evaluation: Method achieves high precision and F1-score on multiple datasets.
- Since the method is unsupervised, it does not require any manual work and can thus be applied broadly to all such application problems
- This work further aids Regology's machine learning framework:

1. Successfully built a keyword extraction algorithm;
2. Increase the performance of existing information retrieval components;
3. Extract topics.

Finding Contextually Consistent Information Units in Legal Text

Dominic Seyler (dseyler2@Illinois.edu)

ChengXiang Zhai (czhai@Illinois.edu)

University of Illinois at Urbana-Champaign

Paul Bruin (paul.bruin@regology.com)

Pavan Bayyapu (pavan.bayyapu@regology.com)

Regology

more info at: https://dominicseyler.com

REGOLOGY ILLINOIS