Mining Information from Statutory Texts in Multi-jurisdictional Settings

Jaromir Savelka
Matthias Grabmair
Kevin D. Ashley

Intelligent Systems Program
University of Pittsburgh

jas438@pitt.edu

JURIX 2014
December 12, 2014
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
Ebola Patient in Texas Presbyterian Hospital
Example Network

Isolated Agents:
- Emergency Management
- Faith Institution
- Transit
- Indian Tribe
- Department of Environmental Protection
- Department of Transportation
- All PH Agents
- Special Response Team

Hints:
- Mouse click + drag: drag the canvas or agent
- Shift + click: show explanation of the agent
- Double click on the agent or connection: open the legal database and show relevant statutes

Keys:
- Arrow up/down: zoom in/out
- L: open/close layout and data menu
- S/P: save a screen snapshot in png/pdf format
- F: freeze/unfreeze the network
- 2: toggle fisheye view
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
1. Set of candidate statutory texts is retrieved on basis of predefined set of search queries from legal IR system.

2. Expert human annotators go through texts and identify relevant spans, i.e. parts containing relevant legal norms.

3. Each relevant span is represented as numeric code following guidelines provided in codebook (citation and 9 descriptors).[27]

NOTE: 95% confidence interval for average inter-annotator agreement for all tasks was reported as (63.1%, 74.9%).

[27] PHASYS Codebook [online]
Example Code Assignment

Example statutory provision

The number of patients admitted to any area of the hospital shall not exceed the number for which the area is designed, equipped, and staffed except in cases of emergency, and then only in accordance with the emergency or disaster plan of the hospital. (28 Pa. Code para 101.172)

Corresponding code

28 Pa. Code § 101.172; Hospital (14); Must Do (2); Suspend (29); Rule/Regulations/Restrictions (4); For Emergency Response (2); Non-specified Disaster/Emergency (5); Public/Individuals (27); Silent (0); Silent (0)
Citation
Relevance
Acting PHS agent (Who is acting?)
Prescription
Action (Which action is being taken?)
Goal
Purpose (For what purpose is action being taken?)
Type of Emergency Disaster
Receiving PHS agent
Timeframe (In what timeframe can/must action be taken?)
Condition

In our work we perform described tasks automatically, i.e.:

1. We transform textual data into feature vectors.
2. We classify vectors in terms of relevance for PHS analysis.
3. We classify vectors in terms of each of nine code categories.
4. We evaluate performance of our system with respect to labels created by expert annotators (treated as gold standard).

In prior work data sparsity was recognized as key element limiting performance.

We focus on use of data from other jurisdictions as one possible way to mitigate problem of data sparsity.
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
Related Work


   We classify texts as containing, e.g., obligation (‘must’), permission (‘may’) or prohibition (‘must not’).

2. Classification of legal literature and legislative texts with hierarchically organized topics.\cite{Francesconi2008a,Opsomer2009}

   Closely related to classification of the texts in terms of relevance.


   We mine texts for presence of similar elements.


   Close to mining texts for specific topical and functional information.

Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
### Comparison of Similar PA and FL Provisions

<table>
<thead>
<tr>
<th>35 P.S. § P.S. 7110.502(1)</th>
<th>Fla. Stat. § 262.60(3)</th>
</tr>
</thead>
</table>

**HEALTH AND SAFETY**

**RADIATION PROTECTION ACT**

**RADIATION EMERGENCY RESPONSE PROGRAM**

Response Program

In conjunction with the department, the agency shall develop a Radiation Emergency Response Program for incorporation into the Pennsylvania Emergency Management Plan developed by the agency pursuant to Title 35 of the Pennsylvania Consolidated Statutes (relating to health and safety). Any volunteer organizations which are incorporated into the Radiation Emergency Response Program developed under the authority of this act shall be consulted prior to such incorporation. The Radiation Emergency Response Program shall include an assessment of potential nuclear accidents or incidents, the radiological consequences and necessary protective measures required to mitigate the effects of such accidents or incidents. The program shall include, but not be limited to:

- Development of a detailed fixed nuclear emergency response plan for areas surrounding each nuclear electrical generation facility, nuclear fuel fabricator and away-from-reactor storage facility. The term "areas" shall be deemed to mean the emergency response zone designated by the NRC Emergency Response Plan applicable to each such fixed nuclear facility.


**MILITARY AFFAIRS AND RELATED MATTERS**

**EMERGENCY MANAGEMENT**

**GENERAL PROVISIONS**

**Radiological** emergency preparedness

Emergency response plans. –In addition to the other plans required by this chapter, the division shall develop, prepare, test, and implement as needed, in conjunction with the appropriate counties and the affected operator, such radiological emergency response plans and preparedness requirements as may be imposed by the United States Nuclear Regulatory Commission or the Federal Emergency Management Agency as a requirement for obtaining or continuing the appropriate licenses for a commercial nuclear electric generating facility.

Differences in Statutory Texts from Different Jurisdictions

Statutory texts coming from different jurisdictions may vary in many aspects:

▶ Different terminology may be used.
▶ Documents may follow different structural patterns.

Even provisions with same purpose may be completely different.

In our experiments we have also observed following differences:

▶ Statutory texts from FL are on average much longer than statutory texts from PA.
▶ Statutory texts from FL are on average more fragmented than statutory texts from PA.

It is difficult to anticipate if performance of classifiers for one state improves by enriching training set with data from another state.
Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
§ 328.307. Licensing of eligible organizations to conduct games of chance

(a) LICENSE REQUIRED.— No eligible organization shall conduct or operate any games of chance unless such eligible organization has obtained and maintains a valid license or limited occasion license issued pursuant to this section. An auxiliary group of a licensed eligible organization shall be eligible to conduct games of chance using the license issued to the eligible organization provided that the auxiliary group or groups are listed on the application and license of the eligible organization. An auxiliary group is not eligible to obtain a license or a limited occasion license. No additional licensing fee shall be charged for an auxiliary group’s eligibility under this chapter. Auxiliary groups shall not include branches, lodges or chapters of a Statewide organization.

(b) ISSUANCE AND FEES.— The licensing authority shall license, upon application, within 30 days any eligible organization meeting the requirements for licensure contained in this chapter to conduct and operate games of chance at such locations within the county or in such manner as stated on the application as limited by subsection (b.1). The license fee to be charged to each eligible organization shall be $100, except for limited occasion licenses which shall be $10. Licenses shall be renewable annually upon the anniversary of the date of issue. The license fee shall be used by the licensing authority to administer this act.

(B.1) LOCATION OF GAMES OF CHANCE.—

(1) Except as otherwise provided in this section, a licensed eligible organization, except a limited occasion licensee, may conduct small games of chance at a licensed premise. The licensed premises shall be...
Partitioning into Subtrees

- Statutory documents are (in comparison to other types of documents) well structured.
- Document can be viewed as a tree graph with given spans of text as nodes and sub-part relations as edges.
- We need to divide each statutory text into smaller parts that could be referred via citations.

(Root) ........................................
   (1) ........................................
      (a) .................................
      (b) .................................
      (c) .................................
   (2) ........................................
§101.62. Request for absentee ballots

(1) (a) The supervisor shall accept a request for an absentee ballot from an elector in person or in writing. One request shall be deemed sufficient to receive an absentee ballot for all elections through the next regularly scheduled general election, unless the elector or the elector’s designee indicates at the time the request is made the elections for which the elector desires to receive an absentee ballot. Such request may be considered canceled when any first-class mail sent by the supervisor to the elector is returned as undeliverable.

(b) The supervisor may accept a written or telephonic request for an absentee ballot from the elector, or, if directly instructed by the elector, a member of the elector’s immediate family, or the elector’s legal guardian. For purposes of this section, the term "immediate family" has the same meaning as specified in paragraph (4)(b). The person making the request must disclose:
1. The name of the elector for whom the ballot is requested.
2. The elector’s address.
3. The elector’s date of birth.
4. The requester’s name.
5. The requester’s address.
6. The requester’s driver’s license number, if available.
7. The requester’s relationship to the elector.
Request for absentee ballots

(1) (a) The supervisor shall accept a request for an absentee ballot from an elector in person or in writing. One request shall be deemed sufficient to receive an absentee ballot for all elections through the next regularly scheduled general election, unless the elector or the elector’s designee indicates at the time the request is made the elections for which the elector desires to receive an absentee ballot. Such request may be considered canceled when any first-class mail sent by the supervisor to the elector is returned as undeliverable.

(b) The supervisor may accept a written or telephonic request for an absentee ballot from the elector, or, if directly instructed by the elector, a member of the elector’s immediate family, or the elector’s legal guardian. For purposes of this section, the term "immediate family" has the same meaning as specified in paragraph (4)(b). The person making the request must disclose:
1. The name of the elector for whom the ballot is requested.
2. The elector’s address.
3. The elector’s date of birth.
4. The requester’s name.
5. The requester’s address.
6. The requester’s driver’s license number, if available.
7. The requester’s relationship to the elector.
Request for absentee ballots

(a) The supervisor shall accept a request for an absentee ballot from an elector in person or in writing. One request shall be deemed sufficient to receive an absentee ballot for all elections through the next regularly scheduled general election, unless the elector or the elector's designee indicates at the time the request is made the elections for which the elector desires to receive an absentee ballot. Such request may be considered canceled when any first-class mail sent by the supervisor to the elector is returned as undeliverable.

(b) The supervisor may accept a written or telephonic request for an absentee ballot from the elector, or, if directly instructed by the elector, a member of the elector's immediate family, or the elector's legal guardian. For purposes of this section, the term "immediate family" has the same meaning as specified in paragraph (4)(b). The person making the request must disclose:
1. The name of the elector for whom the ballot is requested.
2. The elector's address.
3. The elector's date of birth.
4. The requester's name.
5. The requester's address.
6. The requester's driver's license number, if available.
7. The requester's relationship to the elector.
Section 101.62 (2010)

§ 101.62. Request for absentee ballots

(a) The supervisor shall accept a request for an absentee ballot from an elector in person or in writing. One request shall be deemed sufficient to receive an absentee ballot for all elections through the next regularly scheduled general election, unless the elector or the elector’s designee indicates at the time the request is made the elections for which the elector desires to receive an absentee ballot. Such request may be considered canceled when any first-class mail sent by the supervisor to the elector is returned as undeliverable.

(b) The supervisor may accept a written or telephonic request for an absentee ballot from the elector, or, if directly instructed by the elector, a member of the elector’s immediate family, or the elector’s legal guardian. For purposes of this section, the term “immediate family” has the same meaning as specified in paragraph (4)(b). The person making the request must disclose:

1. The name of the elector for whom the ballot is requested.
2. The elector’s address.
3. The elector’s date of birth.
4. The requester’s name.
5. The requester’s address.
6. The requester’s driver’s license number, if available.
7. The requester’s relationship to the elector.

The supervisor may accept a written or telephonic request for an absentee ballot from the elector, or, if directly instructed by the elector, a member of the elector’s immediate family, or the elector’s legal guardian. For purposes of this section, the term “immediate family” has the same meaning as specified in paragraph (4)(b). The person making the request must disclose:

1. The name of the elector for whom the ballot is requested.
2. The elector’s address.
3. The elector’s date of birth.
4. The requester’s name.
5. The requester’s address.
6. The requester’s driver’s license number, if available.
7. The requester’s relationship to the elector.
Request for absentee ballots

The supervisor may accept a written or telephonic request for an absentee ballot from the elector, or, if directly instructed by the elector, a member of the elector’s immediate family, or the elector’s legal guardian. For purposes of this section, the term "immediate family" has the same meaning as specified in paragraph (4)(b). The person making the request must disclose:

The requester’s address.
Corpora (Features)

Pennsylvania

- There were 643 statutory documents to start with (3.02 MB in plain text).
- From these 6022 subtrees were generated.
- Of these 1482 were annotated as relevant by expert human annotators.
- Corpus contains 4765 unique lemmas that were not filtered.

Florida

- There were 399 statutory documents to start with (4.27 MB in plain text).
- From these 11131 subtrees were generated.
- Of these 855 were annotated as relevant by expert human annotators.
- Corpus contains 6570 unique lemmas that were not filtered.
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
In prior work we have shown that NLP techniques and supervised ML approach outperform simple baselines such as:

1. most frequent class (MFC); or

Here we use results obtained from classification performed within single jurisdiction as new baselines.

Aim of our investigation lies exclusively in assessment of effects emulated by use of data from another jurisdiction.

Use of single and simple type of classification model enables us to focus our attention exclusively on effects of adding data.
Data Split into Training and Test Set

We randomly sample whole documents from selected state. Each document has probability 0.2 to be included in sample.

Text units from documents in sample become test set.

Intra-jurisdictional Experiment
Text units from remaining documents from same state become training set.

Cross-jurisdictional Experiment
Text units from all documents from other state become training set.

Multi-jurisdictional Experiment
Text units from remaining documents from same state as well as text units from all documents from other state become training set.
Data Split into Training and Test Set

Intra-jurisdictional  Cross-jurisdictional  Multi-jurisdictional
Different Tasks

There are ten experimental tasks consisting of two-step classification of generated text units.

First step

1. relevance

Second step

2. acting agent
3. prescription
4. action
5. goal
6. purpose
7. emergency type
8. receiving agent
9. time frame
10. condition
Training and Test Set Vectorization

We create vectorized data sets $X^{n \times m}$ with rows as documents and columns as terms by setting each entry of matrix to:

$$
weight(t, d, D) = tf(t, d) \times \log(idf(t, D))
$$

$t$: term
$d$: document
$D$: document collection
$tf(t, d)$: number of occurrences of $t$ in $d$
$idf(t, D)$: number of $d \in D$ over number of $d \in D$ containing $t$

Each $x^{(i)} \in X^{n \times m}$ is vector with $m$ dimensions, where $m$ is number of unique terms that occur in document collection.

Each $x^{(i)} \in X^{n \times m}$ is referenced with unique citation connecting vector to text unit from which it originates.
Training and Test Set Vectorization

Normalized vectors become features: $X_{\text{train}}$ and $X_{\text{test}}$

Depending on task we use respective parts of manually created codebook codes as *labels*: $Y_{\text{train}}$ and $Y_{\text{test}}$

This gives us following *data sets*:

$D_{\text{train}} = (X_{\text{train}}, Y_{\text{train}}) = \{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)}) \}$

$D_{\text{test}} = (X_{\text{test}}, Y_{\text{test}}) = \{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)}) \}$

In each experiment we use respective $D_{\text{train}}$ to train predictive functions $\{ f_1(\cdot), f_2(\cdot), \ldots, f_{10}(\cdot) \}$ for classification of $X_{\text{test}}$.

Classification of each $x^{(i)} \in X_{\text{test}}$ is *automatized counterpart* to manual assessment of corresponding text unit by expert annotator.
Classification Model (tree): CART

Tree is grown from root node by repeatedly using following steps on each node:

1. **Find each predictor’s best split.**
   For each predictor examine each possible subset of values to find best split.

2. **Find node’s best split.** Among best splits found in step 1, choose one that maximizes splitting criterion.

3. **If node is not pure split node.** Split node using its best split found in step 2.

Classification Tree Interpretation

emergeny

w(emergency) < v

0

w(emergency) ≥ v

health

w(health) < v

0

w(health) ≥ v

1
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
Evaluation Metrics

**Precision**
Ratio of correctly classified positive instances over all instances classified as positive.

**Recall**
Ratio of correctly classified positive instances over all true positive instances.

**F₁ Measure**
Harmonic mean of precision and recall where both measures are treated as equally important.

**Accuracy**
Ratio of correctly classified instances over all instances.
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
## Improvement in Multi-jurisdictional Experiment

<table>
<thead>
<tr>
<th></th>
<th>PA→PA</th>
<th>FL→PA</th>
<th>FL+PA→PA</th>
<th>FL→FL</th>
<th>PA→FL</th>
<th>FL+PA→FL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relevance</strong></td>
<td>F: 0.72</td>
<td>F: 0.54</td>
<td>F: 0.73</td>
<td>F: 0.52</td>
<td>F: 0.35</td>
<td>F: 0.54</td>
</tr>
<tr>
<td></td>
<td>P: 0.75</td>
<td>P: 0.62</td>
<td>P: 0.77</td>
<td>P: 0.62</td>
<td>P: 0.27</td>
<td>P: 0.55</td>
</tr>
<tr>
<td></td>
<td>R: 0.70</td>
<td>R: 0.47</td>
<td>R: 0.69</td>
<td>R: 0.45</td>
<td>R: 0.50</td>
<td>R: 0.52</td>
</tr>
<tr>
<td><strong>Act. agent</strong></td>
<td>A: 0.49</td>
<td>A: 0.30</td>
<td>A: 0.52</td>
<td>A: 0.36</td>
<td>A: 0.25</td>
<td>A: 0.44</td>
</tr>
<tr>
<td><strong>Prescription</strong></td>
<td>A: 0.76</td>
<td>A: 0.72</td>
<td>A: 0.77</td>
<td>A: 0.77</td>
<td>A: 0.75</td>
<td>A: 0.75</td>
</tr>
<tr>
<td><strong>Action</strong></td>
<td>A: 0.29</td>
<td>A: 0.23</td>
<td>A: 0.30</td>
<td>A: 0.23</td>
<td>A: 0.18</td>
<td>A: 0.24</td>
</tr>
<tr>
<td><strong>Goal</strong></td>
<td>A: 0.32</td>
<td>A: 0.17</td>
<td>A: 0.32</td>
<td>A: 0.20</td>
<td>A: 0.16</td>
<td>A: 0.25</td>
</tr>
<tr>
<td><strong>Purpose</strong></td>
<td>A: 0.59</td>
<td>A: 0.53</td>
<td>A: 0.61</td>
<td>A: 0.58</td>
<td>A: 0.61</td>
<td>A: 0.62</td>
</tr>
<tr>
<td><strong>Emg. Type</strong></td>
<td>A: 0.78</td>
<td>A: 0.69</td>
<td>A: 0.80</td>
<td>A: 0.76</td>
<td>A: 0.72</td>
<td>A: 0.77</td>
</tr>
<tr>
<td><strong>Rec. agent</strong></td>
<td>A: 0.36</td>
<td>A: 0.25</td>
<td>A: 0.35</td>
<td>A: 0.25</td>
<td>A: 0.25</td>
<td>A: 0.28</td>
</tr>
<tr>
<td><strong>Time frame</strong></td>
<td>A: 0.84</td>
<td>A: 0.81</td>
<td>A: 0.85</td>
<td>A: 0.80</td>
<td>A: 0.78</td>
<td>A: 0.80</td>
</tr>
<tr>
<td><strong>Condition</strong></td>
<td>A: 0.77</td>
<td>A: 0.68</td>
<td>A: 0.75</td>
<td>A: 0.65</td>
<td>A: 0.65</td>
<td>A: 0.67</td>
</tr>
</tbody>
</table>

![Box plots for various categories with different relevance scores and actuator agent statistics.](chart.png)
Similar Traits in Both Jurisdictions

Intra-jurisdictional classifiers trained for Florida (yellow) and Pennsylvania (blue) show that they both share similar traits.
Extendability of the Results

- Results are **promising** because they raise interesting question: Can we improve performance with data from other states?
- Possibly, approach may be useful in other federations or countries that are parts of various international treaties.
- For example, EU directives provide common standards to which member states must adhere with national regulation.
- One can easily imagine context that would call for similar analysis as is here performed for PHS emergency preparedness.
- Results of this work could offer useful insights, although multi-lingual settings would present another challenge.
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
Future Work (General)

▶ Process data from even more states to investigate if performance of classifiers would improve further.

▶ Experiment with techniques to handle imbalanced and sparse data sets, e.g. SMOTE. [6] Chawla et al. 2002


▶ Generate richer text representation (parsing).


▶ Experiment with transfer learning techniques. [19] Pan & Yang 2010
Future Work (Knowledge Based)

- Generate richer text representation (automatic annotators).
- Utilize existing knowledge:
  - codebook\textsuperscript{[27]} Codebook [online]
  - tables of corresponding agents from different states
  - data generated by network analysis
- Integrate rule-based classifiers and feature generators (regular expressions) into existing framework.
Presentation Overview

Motivation

Task Description

Related and Prior Work

Data from Multiple Jurisdictions
  Background and Intuition
  Data Processing
  Experimental Setup
  Evaluation
  Results and Discussion

Future Work

Conclusions
Conclusions

▶ We have presented new results from ongoing effort to use NLP and supervised ML techniques to mine statutory texts.
▶ Performance of most classifiers trained for one state improves if training set is enriched with data from another state.
▶ Possible way to deal with data sparsity was explored and confirmed as promising.
▶ Despite differences in statutory texts from multiple jurisdictions approach leads to improved performance.
References I


Thank you!

Questions, comments and suggestions are welcome now or any time at jas438@pitt.edu.

This work was supported by the University of Pittsburgh's University Research Council Multidisciplinary Small Grant Program. This publication was also supported in part by the Cooperative Agreement 5P01TP000304 from the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CDC.