Information Extraction in Illicit Web Domains

Date: 2017/05/09
Author: Mayank Kejriwal, Pedro Szekely
Source: ACM WWW’ 17
Advisor: Jia-ling Koh
Speaker: Yi-hui Lee
Outline

- Introduction
- Approach
- Experiment
- Conclusion
Introduction

- Information Extraction:
Introduction (cont.)

- Information Extraction on Dark web (human trafficking):

- ages (of human trafficking victims)
- locations
- prices of services
- posting dates
Introduction (cont.)

- A high-level overview of the proposed information extraction approach:

input: Dark Web

output: Annotated corpus
Outline

• Introduction
• Approach
• Experiment
• Conclusion
Approach

• Step 1. Preprocessing:

  - **Readability Text Extractor (RTE):**
    - \(\rightarrow\) **Mercury Web Parser**

  - **NLTK:**
    - RTE string output \(\rightarrow\) sentence tokenize \(\rightarrow\) word tokenize
      - \(\rightarrow\) list of tokens
Approach (cont.)

- Step 2. Apply recognizers:
  - GeoNames-Cities
  - GeoNames-States
  - RegEx-Ages: use regular expressions
  - Dictionary-Names: person names
Approach (cont.)

- Step 3. Word Representation learning:

D1: The cow is in the farm.
D2: I jumped over the farm.
D3: I saw a cow in the farm.

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cow</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>jumped</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>over</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>moon</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>farm</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\text{sim(cow, farm)} = \frac{2}{\sqrt{2} + \sqrt{3}} = 0.64
\]
\[
\text{sim(cow, moon)} = 0
\]
Step 3. Word Representation learning:

- **Random Index [27]**
  - randomly assigned -1, 0, 1 to the vector's attribute
• Step 4. Supervised Contextual Classifier:
  
  Aggregate vectors -> l2-normalization

I saw a cow jumped over the farm

saw = [1, 0, 0, ..., 1, 0]  
a = [1, 1, 1, ..., 1, 1]  
cow = [1, 0, 1, ..., 0, 0]  
jumped = [0, 0, 0, ..., 1, 1]  
over = [0, 0, 1, ..., 0, 1]  
aggregate = [1, 0, 0, ..., 1, 0, 1, 1, 1, ..., 1, 1, ......., 0, 1]  
l2-normalization =  
[0.0001, 0, 0, ..., 0.0001, 0, 0.0001, 0.0001, 0.0001, ..., 0.0001, 0.0001, ......., 0.0000, 1]
Approach (cont.)

- Step 4. Supervised Contextual Classifier:
  - Classifier: Random forest
Outline

• Introduction

• Approach

• Experiment

• Conclusion
Experiment

• Datasets and Ground-truths:
  - Research conducted in the DARPA MEMEX program

Table 2: Four human trafficking corpora for which word representations are (independently) learned

<table>
<thead>
<tr>
<th>Name</th>
<th>Num. websites</th>
<th>Total word count</th>
<th>Unique word count</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-10K</td>
<td>10,000</td>
<td>2,351,036</td>
<td>1,030,469</td>
</tr>
<tr>
<td>D-50K</td>
<td>50,000</td>
<td>11,758,647</td>
<td>5,141,375</td>
</tr>
<tr>
<td>D-100K</td>
<td>100,000</td>
<td>23,536,935</td>
<td>10,277,732</td>
</tr>
<tr>
<td>D-ALL</td>
<td>184,132</td>
<td>43,342,278</td>
<td>18,940,260</td>
</tr>
</tbody>
</table>

• Ground-truths:

Table 3: Five ground-truth datasets on which the classifier (Section 3.4) and baselines are evaluated

<table>
<thead>
<tr>
<th>Name</th>
<th>Pos. ann.</th>
<th>Neg. ann.</th>
<th>Recognizer Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT-Text-City</td>
<td>353</td>
<td>15,783</td>
<td>GeoNames-Cities</td>
</tr>
<tr>
<td>GT-Text-State</td>
<td>100</td>
<td>16,036</td>
<td>GeoNames-States</td>
</tr>
<tr>
<td>GT-Title-City</td>
<td>37</td>
<td>513</td>
<td>GeoNames-Cities</td>
</tr>
<tr>
<td>GT-Text-Name</td>
<td>162</td>
<td>14,337</td>
<td>Dictionary-Names</td>
</tr>
<tr>
<td>GT-Text-Age</td>
<td>116</td>
<td>14,306</td>
<td>RegEx-Ages</td>
</tr>
</tbody>
</table>
Experiment (cont.)

• Baselines:
  
  - Stanford Named Entity Recognition system (NER)
• Evaluation:

Table 5: Comparative results of three systems on precision (P), recall (R) and F1-Measure (F) when training percentage is 30. For the pre-trained baselines, we only report the best results across all applicable models.

<table>
<thead>
<tr>
<th>Ground-truth Dataset</th>
<th>Our System (P/R/F)</th>
<th>Re-trained Baseline (P/R/F)</th>
<th>Pre-trained Baseline (P/R/F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT-Text-City</td>
<td>0.5207/0.5050/0.5116</td>
<td><strong>0.9855/0.1965/0.3225</strong></td>
<td>0.7206/0.7406 <strong>0.7299</strong></td>
</tr>
<tr>
<td>GT-Text-State</td>
<td><strong>0.7852/0.6887 0.7310</strong></td>
<td>0.64/0.0598/0.1032</td>
<td>0.2602/0.8831/0.3993</td>
</tr>
<tr>
<td>GT-Title-City</td>
<td>0.5374/0.5524/0.5406</td>
<td><strong>0.8633/0.1651/0.2685</strong></td>
<td>0.8524/0.7341 <strong>0.7852</strong></td>
</tr>
<tr>
<td>GT-Text-Name</td>
<td>0.7201/0.5850/0.6388</td>
<td>1/0.2103/0.3351</td>
<td>0/0/0</td>
</tr>
<tr>
<td>GT-Text-Age</td>
<td>0.8993/0.9156/0.9068</td>
<td><strong>0.9102/0.7859/0.8412</strong></td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.6925/0.6493/0.6658</td>
<td><strong>0.8798/0.2835/0.3741</strong></td>
<td>0.4583/0.5895/0.4786</td>
</tr>
</tbody>
</table>

Table 6: Comparative results of three systems when training percentage is 70.

<table>
<thead>
<tr>
<th>Ground-truth Dataset</th>
<th>Our System (P/R/F)</th>
<th>Re-trained Baseline (P/R/F)</th>
<th>Pre-trained Baseline (P/R/F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT-Text-City</td>
<td>0.5633/0.6081/0.5841</td>
<td><strong>0.9434/0.3637/0.5000</strong></td>
<td>0.6893/0.7401/0.7128</td>
</tr>
<tr>
<td>GT-Text-State</td>
<td><strong>0.7916/0.7269 0.7502</strong></td>
<td>0.7833/0.2128/0.2971</td>
<td>0.1661/0.7830/0.2655</td>
</tr>
<tr>
<td>GT-Title-City</td>
<td>0.6403/0.6667/0.6437</td>
<td><strong>0.9417/0.3333/0.4790</strong></td>
<td>0.9133/0.6384/0.7289</td>
</tr>
<tr>
<td>GT-Text-Name</td>
<td>0.7174/0.6818/0.6960</td>
<td>1/0.3747/0.5140</td>
<td>0/0/0</td>
</tr>
<tr>
<td>GT-Text-Age</td>
<td>0.9252/0.9273/0.9251</td>
<td><strong>0.9254/0.8454/0.8804</strong></td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.7276/0.7222/0.7198</td>
<td><strong>0.9188/0.4260/0.5341</strong></td>
<td>0.4422/0.5404/0.4268</td>
</tr>
</tbody>
</table>
Experiment (cont.)

• Feature selection:

Figure 5: Effects of additional feature selection on the *GT-Text-Name* dataset (30% training data)
Outline

• Introduction
• Approach
• Experiment
• Conclusion
Conclusion

• We presented a lightweight, feature-agnostic Information Extraction approach that is suitable for illicit Web domains.

• Our approach relies on unsupervised derivation of word representations from an initial corpus, and the training of a supervised contextual classifier using external high-recall recognizers and a handful of manually verified annotations.

• Real-world settings:

  - End Human Trafficking hackathon organized by the office of the District Attorney of New York17