Using Contexts and Constraints for Improved Geotagging of Human Trafficking Webpages

Rahul Kapoor, Mayank Kejriwal and Pedro Szekely

Information Sciences Institute, USC Viterbi School of Engineering
Domain-specific Insight Graphs (DIG)
Geotagging HT webpages

- Disambiguation problem: is **Charlotte** a name or city? Depends on context!
Geotagging HT webpages

• **Toponym Resolution**

• Examples –
  – “Kansas City” is a city in the state Missouri as well as Kansas
  – “Los Angeles” is also a town in Texas apart from being a city in California
Potential approach: use Geonames

- Open database of geolocations
- Contains 2.8 million populated places in the world along with 5.5 million alternate names
- Each has a unique id and details of the state, country, latitude, longitude, population
- Due to the large size, we use Trie based approach for high recall dictionary extraction

Using Geonames lexicon for extractions

- Common words like “the”, “makes”, “falls” are city names as well
- Some abbreviations used in the text are also marked as cities
Contexts and constraints are both important

• Constraints reflect **domain** knowledge (‘semantics’ of the domain e.g., that a city is in a state and a state is in a country; also, **a priori** knowledge)

• Context reflect statistical (aka **data-driven**) knowledge
Use context to train word embeddings

- Many options in the literature (word2vec, random indexing...)
- Random indexing found to work well for HT in previous work
Useful for assigning probabilities to extractions in 200 dimension vector space.
Context-based classifier
How do we encode constraints?

• By itself, context is not enough; more can be done to improve performance!
• Integer Linear Programming is an established framework
• Requires manual crafting of:
  • Objective functions
  • Linear Constraints
  • Weights
OBJECTIVE FUNCTION WEIGHTS
Token Source Weight

- Captures relative importance of source of extraction
- City appearing in title is more important than those in footer
Context Weight

- Captures what extraction is more likely to be correct depending on the context
- “I am new to Charlotte”, “My name is Charlotte” - in the 1\textsuperscript{st} sentence the same word is more likely to be a city than in 2\textsuperscript{nd}
Population Weight

- Larger cities are more likely to be referred than smaller cities
- When someone mentions “Los Angeles”, he is most likely not referring to a small town in TX but the much larger city in CA
CONSTRAINTS
Semantic Type Exclusivity

- An extraction marked as multiple semantic types can be only one of those.
- Charlotte_City + Charlotte_Name <= 1, means “Charlotte” can be either a city name or a name of a person at a time.

\[ \forall \text{candidate}_i \in \text{candidates}, \sum_{j=1}^{\text{types}} \text{candidate}_{i\text{type}_j} \leq 1 \]
Extractions of a Semantic Type

- Limits the number of extractions of a page
- \( \text{LosAngeles\_City} + \text{Seattle\_City} + \text{Houston\_City} \leq 1 \), means atmost one of the cities can be selected

\[
\forall type_j \in \text{types}, \quad \sum_{i=1}^{\text{candidates}} candidate_{i\_type_j} \leq 1
\]
Valid City–State-Country Combination

- The selected city should be in the selected country/state
- LosAngeles_US + NewYorkCity_US <= US, means if one of the cities on the left is selected, the country on the right must be selected

\[
\forall country_j \in countries, \sum_{i=1}^{\text{cities}} city_i \cdot country_j \leq country_j
\]

\[
\forall state_j \in states, \sum_{i=1}^{\text{cities}} city_i \cdot state_j \leq state_j
\]
City-State/Country Exclusivity

- The chosen city has a corresponding state/country selected
- Portland_Oregon + Portland_Maine = Portland, means if Portland is selected, one of its corresponding states must be selected

\[
\forall city_i \in \text{cities}, \sum_{j=1}^{\text{countries}} city_i \text{country}_j = city_i
\]

\[
\forall city_i \in \text{cities}, \sum_{j=1}^{\text{states}} city_i \text{state}_j = city_i
\]
Putting it together...
EXPERIMENTS
Dataset

• Word Embeddings trained on a corpus of 90,000 web pages, using Random Indexing
• Context classifier trained on 75 webpages
• Groundtruth for ILP contained smaller corpus of 20 webpages coming from 10 different domains, having 175 geolocation annotations
Comparison

- The extractions from ILP are compared to:
  - Random: A random selection from the extractions
  - Top Ranked: The highest ranked extraction according to the context probabilities
- Metrics: Precision, Recall of extractions
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.5</td>
<td>0.35714286</td>
</tr>
<tr>
<td>Top Ranked</td>
<td>0.61538462</td>
<td>0.57142857</td>
</tr>
<tr>
<td>ILP</td>
<td>0.78571429</td>
<td>0.78571429</td>
</tr>
</tbody>
</table>
## Future Work

- Using Probabilistic Soft Logic as an alternative to model the problem

<table>
<thead>
<tr>
<th>ILP</th>
<th>PSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>As the factors affecting selection increase, need to combine weights for objective function</td>
<td>Probabilistic model with continuous random variables allows to capture multiple factors</td>
</tr>
<tr>
<td>Not possible to model complex relations which affect extraction selection</td>
<td>Can model based on First Order Logic representation</td>
</tr>
<tr>
<td>Each extraction is either selected or not selected</td>
<td>Each extraction can be assigned an expectation value</td>
</tr>
<tr>
<td>May take time to optimize</td>
<td>Soft truth values enable faster convergence</td>
</tr>
</tbody>
</table>

Refer: [http://psl.linqs.org/](http://psl.linqs.org/)