Overview

- **Problem**: harvesting information from data sources can be challenging because data is published in different data formats and using different conventions.
- **Goal**: build a semantic model that describes the data source.

**Input**:
- A set of domain ontologies $O$
- A target data source $s(a_1, a_2, ..., a_k)$: $a_i$ is a source attribute

**Output**:
- A semantic model $sm(s)$

**Application**: automatically convert data sources to RDF triples to publish to knowledge graphs.

How to navigate in the search space

- Construct a space of possible semantic models ($G_{int}$)

  $f(s_{int}) = \sum_{\theta} \exp \left( \sum_{\mathbf{c}} \mathbf{w}_c \right) \prod_{i=1}^{k} \mathbf{w}_i \mathbf{x}_i \mathbf{y}_i$

  - Transition function $f$: merge new attribute to existing semantic model according to $G_{int}$

How to rank semantic models

- Likelihood of a semantic model - $P(\forall y \in y; y = true|x)$ - indicates the quality of the model
- Estimate the likelihood using a conditional random field (CRF)
  - $P(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \prod_{\mathbf{w}_c} \frac{\Psi_c(y_c; \theta)}{\Psi_{true}(y_c; \theta)}$
  - $Z(x) = \sum_x \prod_{c \in C} \prod_{\mathbf{w}_c} \Psi_c(y_c, x_c; \theta_p)$

  - $\Psi_{true}(y_c, x_c; \theta_p) = \exp \left[ \sum_{\mathbf{c}} \mathbf{w}_c \mathbf{x}_i \mathbf{y}_i \right]$

Features:
- Link confidence
- Cardinality relationships between source attributes
- Structural similarity
...

Training the Graphical Model

- Create labeled data from sample of possible semantic models
- Train to identify correct/incorrect links in the models

Evaluation

- **Datasets**: museum-crm and museum-edm (Taheriyan 2016)
- **Semantic labelers**: DSL (Pham et al. 2016), SemTyper (Ramnandan et al. 2015), Serene (Rummele et al. 2018)

<table>
<thead>
<tr>
<th>Method</th>
<th>museum-crm</th>
<th>museum-edm</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemTyper</td>
<td>0.830</td>
<td>0.886</td>
</tr>
<tr>
<td>DSL</td>
<td>0.628</td>
<td>0.896</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.912</td>
<td>0.910</td>
</tr>
</tbody>
</table>

Baselines: (Taheriyan et al. 2016), Serene (Una et al. 2018)

Conclusion and Future Work

- By exploiting relationships within the data sources and semantic models, our approach:
  - generates better semantic models
  - is more robust to noise

- Future work:
  - Minimize user effort by leveraging Linked Open Data
  - End-to-end system from web extraction to semantic model
  - Integrate with interactive modeling system (Karma)