LARGE-SCALE KNOWLEDGE GRAPH IDENTIFICATION USING PSL

Jay Pujara¹, Hui Miao¹, Lise Getoor¹, William Cohen²
¹University of Maryland, College Park, US
²Carnegie Mellon University

AAAI Symposium on Semantics for Big Data
11/16/2013
Overview

Problem: Build a Knowledge Graph from millions of noisy extractions

Method: Use probabilistic soft logic to easily specify models and efficiently optimize them

Approach: Knowledge Graph Identification reasons jointly over all facts in the knowledge graph

Results: State-of-the-art performance on real-world datasets producing knowledge graphs with millions of facts
CHALLENGES IN KNOWLEDGE GRAPH CONSTRUCTION
Motivating Problem: New Opportunities

Internet

Massive source of publicly available information

Extraction

Cutting-edge IE methods

Knowledge Graph (KG)

Structured representation of entities, their labels and the relationships between them
Motivating Problem: Real Challenges

- Internet
  - Noisy!

- Extraction
  - Difficult!

- Knowledge Graph
  - Contains many errors and inconsistencies
NELL: The Never-Ending Language Learner

- Large-scale IE project (Carlson et al., 2010)
- Lifelong learning: aims to “read the web”
- Ontology of known labels and relations
- Knowledge base contains millions of facts
Examples of NELL errors
Kyrgyzstan has many variants:
- Kyrgystan
- Kyrgistan
- Kyrghyzstan
- Kyrgzstan
- Kyrgyz Republic

Saudi Cultural Days in the Kyrgyz Republic has concluded its activities in the capital Bishkek in the weekend in a special ceremony held on this occasion. The event was attended by Deputy Minister of Culture and Tourism of the Kyrgyz Republic Koulev Mirza; Kyrgyzstan's Ambassador to Saudi Arabia Jusupbek Sharipov; the Saudi Embassy Acting Chargé d'affaires to Kyrgyzstan, Mari bin Barakah Al-Derbas and members of the embassy staff, in the presence of a heavy turnout of Kyrgyz citizens.

The Days of Culture of Saudi Arabia in Kyrgyzstan will be held from 6 to 9 May.

Refugees are often from areas where conflict is historically embedded and marked in ideology and injustice. The Tsarnaev family emigrated from the Chechen diaspora in Kyrgyzstan, a region Stalin deported the Chechens to in 1943. After the fall of the Berlin Wall in 1991, Chechens engaged in a battle for independence from Russia that led to the Tsarnaevs' petition for refugee status in the early
Kyrgyzstan is labeled a bird and a country.

Anssi Kullberg has sent along some great trip reports to unusual places, including Kyrgyzstan, Pakistan, Egypt/Jordan, and Afghanistan. I had to create a whole new country page for Afghanistan to hold that last one! Thanks so much, Anssi!

Erik Kleyheeg has just returned from Lesvos with some new bird images. Included here are: Common Scops-Owl, Wood Warbler, Spanish Sparrow, Red-throated Pipit, Eurasian Chiff-chaff, and Cretzschmar’s Bunting.

Kyrgyzstan (Kyrgyz: Кыргызстан; IPA: [kɔrɡiˈstɑn]; Russian: Киргизия), officially the Kyrgyz Republic (Кыргыз Республикасы; Russian: Кыргызская Республика), is a country located in Central Asia. Landlocked and mountainous, Kyrgyzstan is bordered by Kazakhstan to the north, Uzbekistan to the west, Tajikistan to the southwest and China to the east. Its capital and largest city is Bishkek.
Missing and spurious relations

Guidance
Kazakhstan / Kyrgyzstan – Consular Fees

 Kyrgyzstan’s location is ambiguous – Kazakhstan, Russia and US are included in possible locations

Kyrgyzstan U.S. Air Base Future Unclear

A Central Asian country of incredible natural beauty and proud nomadic traditions, most of Kyrgyzstan was formally annexed to Russia in 1876. The Kyrgyz staged a major revolt against the Tsarist Empire in 1916 in which almost one-sixth of the Kyrgyz population was killed. Kyrgyzstan became a Soviet republic in 1936 and
Violations of ontological knowledge

- Equivalence of co-referent entities (sameAs)
  - SameEntity(Kyrgyzstan, Kyrgyz Republic)
- Mutual exclusion (disjointWith) of labels
  - MUT(bird, country)
- Selectional preferences (domain/range) of relations
  - RNG(countryLocation, continent)

Enforcing these constraints require jointly considering multiple extractions
Motivating Problem (revised)

Internet

(noisy) Extraction Graph

Large-scale IE

Knowledge Graph

Joint Reasoning
Knowledge Graph Identification

Problem:

Extraction Graph

Knowledge Graph Identification

Solution: Knowledge Graph Identification (KGI)

- Performs graph identification:
  - entity resolution
  - collective classification
  - link prediction
- Enforces ontological constraints
- Incorporates multiple uncertain sources
Illustration of KGI: Extractions

Uncertain Extractions:
.5: Lbl(Kyrgyzstan, bird)
.7: Lbl(Kyrgyzstan, country)
.9: Lbl(Kyrgyz Republic, country)
.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)
**Uncertain Extractions:**
.5: Lbl(Kyrgyzstan, bird)
.7: Lbl(Kyrgyzstan, country)
.9: Lbl(Kyrgyz Republic, country)
.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)
Illustration of KGI: Ontology + ER

**Uncertain Extractions:**
- .5: Lbl(Kyrgyzstan, bird)
- .7: Lbl(Kyrgyzstan, country)
- .9: Lbl(Kyrgyz Republic, country)
- .8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

**Ontology:**
- Dom(hasCapital, country)
- Mut(country, bird)

**Entity Resolution:**
- SameEnt(Kyrgyz Republic, Kyrgyzstan)
Uncertain Extractions:
.5: Lbl(Kyrgyzstan, bird)
.7: Lbl(Kyrgyzstan, country)
.9: Lbl(Kyrgyz Republic, country)
.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

Ontology:
Dom(hasCapital, country)
Mut(country, bird)

Entity Resolution:
SameEnt(Kyrgyz Republic, Kyrgyzstan)

After Knowledge Graph Identification

MODELING KNOWLEDGE
GRAPH IDENTIFICATION
Viewing KGI as a probabilistic graphical model

Lbl(Kyrgyzstan, bird) → Lbl(Kyrgyzstan, country) → Rel(hasCapital, Kyrgyzstan, Bishkek)

Lbl(Kyrgyz Republic, bird) → Lbl(Kyrgyz Republic, country) → Rel(hasCapital, Kyrgyz Republic, Bishkek)
Background: Probabilistic Soft Logic (PSL)

- Templating language for hinge-loss MRFs, very scalable!
- Model specified as a collection of logical formulas

\[
\text{SAMEENT}(E_1, E_2) \overset{\sim}{\implies} \text{LBL}(E_1, L) \implies \text{LBL}(E_2, L)
\]

- Uses soft-logic formulation
  - Truth values of atoms relaxed to [0, 1] interval
  - Truth values of formulas derived from Lukasiewicz t-norm
Background: PSL Rules to Distributions

- Rules are *grounded* by substituting literals into formulas

\[ w_{EL} : \text{SAMEENT}(\text{Kyrgyzstan, Kyrgyz Republic}) \sim \text{LBL}(\text{Kyrgyzstan, country}) \Rightarrow \text{LBL}(\text{Kyrgyz Republic, country}) \]

- Each ground rule has a weighted *distance to satisfaction* derived from the formula’s truth value

\[
P(G \mid E) = \frac{1}{Z} \exp \left[ - \sum_{r \in R} w_r \varphi_r (G) \right]
\]

- The PSL program can be interpreted as a joint probability distribution over all variables in knowledge graph, conditioned on the extractions
Background: Finding the best knowledge graph

• MPE inference solves $\max_G P(G)$ to find the best KG

• In PSL, inference solved by convex optimization

• Efficient: running time scales with $O(|R|)$
PSL Rules for the KGI Model
PSL Rules: Uncertain Extractions

Weight for source T (relations)

\[ w_{CR-T} : \text{CANDREL}_T(E_1, E_2, R) \]

Predicate representing uncertain relation extraction from extractor T

Relation in Knowledge Graph

\[ \Rightarrow \text{REL}(E_1, E_2, R) \]

Weight for source T (labels)

\[ w_{CL-T} : \text{CANDLBL}_T(E, L) \]

Predicate representing uncertain label extraction from extractor T

Label in Knowledge Graph

\[ \Rightarrow \text{LBL}(E, L) \]
PSL Rules: Entity Resolution

\[ w_{EL} : \text{SAMEENT}(E_1, E_2) \land LBL(E_1, L) \Rightarrow LBL(E_2, L) \]
\[ w_{ER} : \text{SAMEENT}(E_1, E_2) \land \text{REL}(E_1, E, R) \Rightarrow \text{REL}(E_2, E, R) \]
\[ w_{ER} : \text{SAMEENT}(E_1, E_2) \land \text{REL}(E, E_1, R) \Rightarrow \text{REL}(E, E_2, R) \]

ER predicate captures confidence that entities are co-referent

• Rules require co-referent entities to have the same labels and relations

• Creates an equivalence class of co-referent entities
PSL Rules: Ontology

Inverse:
\[
\text{wo} : \ \text{INV}(R, S) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{REL}(E_2, E_1, S)
\]

Selectional Preference:
\[
\begin{align*}
\text{wo} : \ \text{DOM}(R, L) & \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{LBL}(E_1, L) \\
\text{wo} : \ \text{RNG}(R, L) & \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{LBL}(E_2, L)
\end{align*}
\]

Subsumption:
\[
\begin{align*}
\text{wo} : \ \text{SUB}(L, P) & \quad \land \quad \text{LBL}(E, L) \quad \Rightarrow \quad \text{LBL}(E, P) \\
\text{wo} : \ \text{RSUB}(R, S) & \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{REL}(E_1, E_2, S)
\end{align*}
\]

Mutual Exclusion:
\[
\begin{align*}
\text{wo} : \ \text{MUT}(L_1, L_2) & \quad \land \quad \text{LBL}(E, L_1) \quad \Rightarrow \quad \neg \text{LBL}(E, L_2) \\
\text{wo} : \ \text{RMUT}(R, S) & \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \neg \text{REL}(E_1, E_2, S)
\end{align*}
\]

Adapted from Jiang et al., ICDM 2012
Probability Distribution over KGs

\[ P(G \mid E) = \frac{1}{Z} \exp \left[ - \sum_{r \in R} w_r \varphi_r (G) \right] \]

- \text{CANDLBL}_T(\text{kyrgyzstan, bird}) \Rightarrow \text{LBL}(\text{kyrgyzstan, bird})
- \text{MUT}(\text{bird, country})
  \Rightarrow \hat{\text{LBL}}(\text{kyrgyzstan, bird})
  \Rightarrow \hat{\neg}\text{LBL}(\text{kyrgyzstan, country})
- \text{SAMEENT}(\text{kyrgz republic, kyrgyzstan})
  \Rightarrow \hat{\text{LBL}}(\text{kyrgz republic, country})
  \Rightarrow \text{LBL}(\text{kyrgyzstan, country})
EVALUATION
# Two Evaluation Datasets

<table>
<thead>
<tr>
<th></th>
<th>LinkedBrainz</th>
<th>NELL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Community-supplied data about musical artists, labels, and creative works</td>
<td>Real-world IE system extracting general facts from the WWW</td>
</tr>
<tr>
<td><strong>Noise</strong></td>
<td>Realistic synthetic noise</td>
<td>Imperfect extractors and ambiguous web pages</td>
</tr>
<tr>
<td><strong>Candidate Facts</strong></td>
<td>810K</td>
<td>1.3M</td>
</tr>
<tr>
<td><strong>Unique Labels and Relations</strong></td>
<td>27</td>
<td>456</td>
</tr>
<tr>
<td><strong>Ontological Constraints</strong></td>
<td>49</td>
<td>67.9K</td>
</tr>
</tbody>
</table>
LinkedBrainz dataset for KGI

Mapping to FRBR/FOAF ontology

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOM</td>
<td>rdfs:domain</td>
</tr>
<tr>
<td>RNG</td>
<td>rdfs:range</td>
</tr>
<tr>
<td>INV</td>
<td>owl:inverseOf</td>
</tr>
<tr>
<td>SUB</td>
<td>rdfs:subClassOf</td>
</tr>
<tr>
<td>RSUB</td>
<td>rdfs:subPropertyOf</td>
</tr>
<tr>
<td>MUT</td>
<td>owl:disjointWith</td>
</tr>
</tbody>
</table>
Adding noise to LinkedBrainz

Add realistic noise to LinkedBrainz data:

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Erroneous Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-reference</td>
<td>User misspells artist</td>
</tr>
<tr>
<td>Label</td>
<td>User swaps artist and album fields</td>
</tr>
<tr>
<td>Relation</td>
<td>User omits or adds spurious albums for artist</td>
</tr>
<tr>
<td>Reliability</td>
<td>Gaussian noise on truth value of information</td>
</tr>
</tbody>
</table>
**LinkedBrainz experiments**

Comparisons:

- **Baseline**: Use noisy truth values as fact scores
- **PSL-EROOnly**: Only apply rules for Entity Resolution
- **PSL-OntOnly**: Only apply rules for Ontological reasoning
- **PSL-KGI**: Apply Knowledge Graph Identification model

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 at .5</th>
<th>Max F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.672</td>
<td>0.946</td>
<td>0.477</td>
<td>0.634</td>
<td>0.788</td>
</tr>
<tr>
<td>PSL-EROOnly</td>
<td>0.797</td>
<td>0.953</td>
<td>0.558</td>
<td>0.703</td>
<td>0.831</td>
</tr>
<tr>
<td>PSL-OntOnly</td>
<td>0.753</td>
<td>0.964</td>
<td>0.605</td>
<td>0.743</td>
<td>0.832</td>
</tr>
<tr>
<td>PSL-KGI</td>
<td>0.901</td>
<td>0.970</td>
<td>0.714</td>
<td>0.823</td>
<td>0.919</td>
</tr>
</tbody>
</table>
NELL Evaluation: two settings

Target Set: restrict to a subset of KG

- Closed-world model
- Uses a target set: subset of KG
- Derived from 2-hop neighborhood
- Excludes trivially satisfied variables

Complete: Infer full knowledge graph

- Open-world model
- All possible entities, relations, labels
- Inference assigns truth value to each variable
NELL experiments:

Target Set

Task: Compute truth values of a target set derived from the evaluation data

Comparisons:

Baseline Average confidences of extractors for each fact in the NELL candidates
NELL Evaluate NELL’s promotions (on the full knowledge graph)
MLN Method of (Jiang, ICDM12) – estimates marginal probabilities with MC-SAT
PSL-KGI Apply full Knowledge Graph Identification model

Running Time: Inference completes in 10 seconds, values for 25K facts

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>.873</td>
<td>.828</td>
</tr>
<tr>
<td>NELL</td>
<td>.765</td>
<td>.673</td>
</tr>
<tr>
<td>MLN (Jiang, 12)</td>
<td>.899</td>
<td>.836</td>
</tr>
<tr>
<td>PSL-KGI</td>
<td>.904</td>
<td>.853</td>
</tr>
</tbody>
</table>
NELL experiments:
Complete knowledge graph

**Task:** Compute a full knowledge graph from uncertain extractions

**Comparisons:**
- **NELL**  
  NELL’s strategy: ensure ontological consistency with existing KB
- **PSL-KGI**  
  Apply full Knowledge Graph Identification model

**Running Time:** Inference completes in 130 minutes, producing 4.3M facts

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NELL</td>
<td>0.765</td>
<td>0.801</td>
<td>0.477</td>
<td>0.634</td>
</tr>
<tr>
<td>PSL-KGI</td>
<td>0.892</td>
<td>0.826</td>
<td>0.871</td>
<td>0.848</td>
</tr>
</tbody>
</table>
Conclusion

• Knowledge Graph Identification is a powerful technique for producing knowledge graphs from noisy IE system output

• Using PSL we are able to enforce global ontological constraints and capture uncertainty in our model

• Unlike previous work, our approach infers complete knowledge graphs for datasets with millions of extractions

Code available on GitHub:

https://github.com/linqs/KnowledgeGraphIdentification

Questions?