Can Computers Create Knowledge?

Internet

Massive source of publicly available information

Knowledge
Computers + Knowledge =❤️
Motivating Problem: New Opportunities

Internet

Massive source of publicly available information

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
Knowledge Graphs in the wild
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
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

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
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 (/kərˈɡɪstən/ kur-ɡɪ-stən;[5] Kyrgyz: Кыргызстан (IPA: [qyrɡɪsˈstan])); Russian: Киргизия), officially the Kyrgyz Republic (Kyrgyz: Кыргыз Республикасы; Russian: Кыргызская Республика), is a country located in Central Asia.[6] 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.
Kyrgyzstan’s location is ambiguous – Kazakhstan, Russia and US are included in possible locations.
Violations of ontological knowledge

• Equivalence of co-referent entities (sameAs)
  • SameAs(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
KNOWLEDGE GRAPH IDENTIFICATION
Motivating Problem (revised)

Internet \rightarrow \text{(noisy) Extraction Graph} \rightarrow \text{Large-scale IE} \rightarrow \text{Knowledge Graph} = \text{Joint Reasoning}
Knowledge Graph Identification

**Problem:**
- Extraction Graph

**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)
Illustration of KGI: Extraction Graph

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)
Illustration of KGI

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)

(Annotated) Extraction Graph

After Knowledge Graph Identification
Viewing KGI as a probabilistic graphical model

Lbl(Kyrgyzstan, country)

Lbl(Kyrgyzstan, bird)

Lbl(Kyrgyz Republic, country)

Rel(hasCapital, Kyrgyzstan, Bishkek)

Rel(hasCapital, Kyrgyz Republic, Bishkek)

Lbl(Kyrgyz Republic, bird)

(Pujara et al., ISWC13)
Background: Probabilistic Soft Logic (PSL)

(Broecheler et al., UAI10; Kimming et al., NIPS-ProbProg12)

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

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

- 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

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

• 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

(Pujara et al., ISWC13)
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 empirically scales with $O(|R|)$
  (Bach et al., NIPS12)
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

Predicate representing uncertain label extraction from extractor T
\( w_{CL-T} : \text{CANDLBL}_T(E, L) \)

Relation in Knowledge Graph
\( \Rightarrow \text{REL}(E_1, E_2, R) \)

Label in Knowledge Graph
\( \Rightarrow \text{LBL}(E, L) \)

(Pujara et al., ISWC13)
PSL Rules: Entity Resolution

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

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

SameEnt 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:

\[ w_0 : \text{INV}(R, S) \; \land \; \text{REL}(E_1, E_2, R) \implies \text{REL}(E_2, E_1, S) \]

Selectional Preference:

\[ w_0 : \text{DOM}(R, L) \; \land \; \text{REL}(E_1, E_2, R) \implies \text{LBL}(E_1, L) \]
\[ w_0 : \text{RNG}(R, L) \; \land \; \text{REL}(E_1, E_2, R) \implies \text{LBL}(E_2, L) \]

Subsumption:

\[ w_0 : \text{SUB}(L, P) \; \land \; \text{LBL}(E, L) \implies \text{LBL}(E, P) \]
\[ w_0 : \text{RSUB}(R, S) \; \land \; \text{REL}(E_1, E_2, R) \implies \neg \text{REL}(E_1, E_2, S) \]

Mutual Exclusion:

\[ w_0 : \text{MUT}(L_1, L_2) \; \land \; \text{LBL}(E, L_1) \implies \neg \text{LBL}(E, L_2) \]
\[ w_0 : \text{RMUT}(R, S) \; \land \; \text{REL}(E_1, E_2, R) \implies \neg \text{REL}(E_1, E_2, S) \]

Adapted from Jiang et al., ICDM 2012
\( \phi_1 \) \text{CANDLBL}\text{struct}(Kyrgyzstan, bird) \\
\Rightarrow \text{LBL}(Kyrgyzstan, bird)

\( \phi_2 \) \text{CANDREL}\text{pat}(Kyrgyz Rep., Asia, locatedIn) \\
\Rightarrow \text{REL}(Kyrgyz Rep., Asia, locatedIn)

\( \phi_3 \) \text{SAMEENT}(Kyrgyz Rep., Kyrgyzstan) \\
\land \text{LBL}(Kyrgyz Rep., country) \\
\Rightarrow \text{LBL}(Kyrgyzstan, country)

\( \phi_4 \) \text{DOM}(locatedIn, country) \\
\land \text{REL}(Kyrgyz Rep., Asia, locatedIn) \\
\Rightarrow \text{LBL}(Kyrgyz Rep., country)

\( \phi_5 \) \text{MUT}(country, bird) \\
\land \text{LBL}(Kyrgyzstan, country) \\
\Rightarrow \neg \text{LBL}(Kyrgyzstan, bird)
Probability Distribution over KGs

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

- \text{CANDLBT}(\text{kyrgyzstan}, \text{bird}) \Rightarrow \text{LBL}(\text{kyrgyzstan}, \text{bird})
- \text{MUT} (\text{bird}, \text{country})
  \Rightarrow \neg \text{LBL}(\text{country})
- \text{SAMEENT}(\text{kyrgz republic}, \text{kyrgyzstan})
  \Rightarrow \neg \text{LBL}(\text{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,</td>
<td>Real-world IE system extracting general</td>
</tr>
<tr>
<td></td>
<td>labels, and creative works</td>
<td>facts from the WWW</td>
</tr>
<tr>
<td><strong>Noise</strong></td>
<td>Realistic synthetic noise</td>
<td>Imperfect extractors and ambiguous web</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pages</td>
</tr>
<tr>
<td><strong>Candidate Facts</strong></td>
<td>810K</td>
<td>1.3M</td>
</tr>
<tr>
<td><strong>Unique Labels</strong></td>
<td>27</td>
<td>456</td>
</tr>
<tr>
<td>and Relations</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ontological</strong></td>
<td>49</td>
<td>67.9K</td>
</tr>
<tr>
<td><strong>Constraints</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Pujara et al., ISWC13)
LinkedBrainz

- Open source community-driven structured database of music metadata
- Uses proprietary schema to represent data

LinkedBrainz project provides an RDF mapping from MusicBrainz data to Music Ontology using the D2RQ tool

- Built on popular ontologies such as FOAF and FRBR
- Widely used for music data (e.g. BBC Music Site)

(Pujara et al., ISWC13)
LinkedBrainz dataset for KGI

Mapping to FRBR/FOAF ontology

<table>
<thead>
<tr>
<th>DOM</th>
<th>rdfs:domain</th>
</tr>
</thead>
<tbody>
<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>

(Pujara et al., ISWC13)
## 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><strong>0.901</strong></td>
<td><strong>0.970</strong></td>
<td><strong>0.714</strong></td>
<td><strong>0.823</strong></td>
<td><strong>0.919</strong></td>
</tr>
</tbody>
</table>

(Pujara et al., ISWC13)
NELL Evaluation: two settings

Target Set: restrict to a subset of KG
(Jiang, ICDM12)

- 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>F1</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:**

<table>
<thead>
<tr>
<th></th>
<th>NELL’s strategy: ensure ontological consistency with existing KB</th>
<th>PSL-KGI: Apply full Knowledge Graph Identification model</th>
</tr>
</thead>
</table>

**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>

(Pujara et al., ISWC13)
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?
Problem: Incremental Updates to KG

How do we add new extractions to the Knowledge Graph?
Naïve Approach: Full KGI over extractions
Improving the naïve approach

**Intuition:** Much of previous KG does not change

**Online collective inference:**
- Selectively update the MAP state
- Bound the *regret* of partial updates
- Efficiently determine which variables to infer
Key Idea: fix some variables, infer others
Key Collaborators
Approximation: KGI over subset of graph
Theory: Bounding Inference Regret

Regret = \|\text{full inference} - \text{partial update}\|
Theory: Bounding Inference Regret

\[ \text{Regret} = \| \text{full inference} - \text{partial update} \| \]

\[ R_n(x, y_S; \dot{w}) \triangleq \frac{1}{n} \| h(x; \dot{w}) - h(x, y_S; \dot{w}) \|_1 \]
Theory: Bounding Inference Regret

\[ \mathcal{R}_n(x, y_S; \hat{\omega}) \triangleq \frac{1}{n} \| h(x; \hat{\omega}) - h(x, y_S; \hat{\omega}) \|_1 \]

\[ \mathcal{R}_n(x, y_S; \hat{\omega}) \leq O \left( \sqrt{\frac{B \| \omega \|_2}{n \cdot w_p}} \| y_S - \hat{y}_S \|_1 \right) \]