Can Computers Create Knowledge?

Internet

Massive source of publicly available information

Knowledge
Computers + Knowledge = ❤️

New York Giants
4-6, 3rd in NFC Eastern Division

Yesterday, 4:25 PM (ET)
MetLife Stadium, East Rutherford, New Jersey

Green Bay Packers (5-6) 13 - 27 Final

New York Giants (4-6)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packers</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Giants</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td>27</td>
</tr>
</tbody>
</table>

Sun, Nov 24 vs. Cowboys 4:25 PM (ET)

News for Giants

People I know who studied at University of Maryland, College Park

Friends of people I know who studied at University of Maryland, College Park

Photos of people I know who studied at University of Maryland, College Park

Photos by people I know who studied at University of Maryland, College Park

“What sort of Pokémon is Pikachu”

The answer is electric.

Input interpretation
- Pikachu type

Result
electric

Basic properties

<table>
<thead>
<tr>
<th>English name</th>
<th>Pikachu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese name</td>
<td>ピカチュウ (Pikachu)</td>
</tr>
<tr>
<td>Pokédex number</td>
<td>25</td>
</tr>
<tr>
<td>type</td>
<td>electric</td>
</tr>
</tbody>
</table>
What does it mean to create knowledge?
What do we mean by knowledge?
Defining the Questions

• Extraction

• Representation

• Reasoning and Inference
Defining the Questions

• Extraction

• Representation

• Reasoning and Inference
A Revised Knowledge-Creation Diagram

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
Knowledge Graphs in the wild
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., AAAI10)

• 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
- Kyrgyzstan
- 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.

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.

Anssi Kullberg has sent along some great trip reports to unusual places, including Kyrgyzstan, Pakistan.

Kyrgyzstan (/kærɡəˈstɑːn/ kur-gə-stahn;[5] Kyrgyz: Кыргызстан (IPA: [kyrɡɨˈstan]); 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.
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 requires jointly considering multiple extractions across documents
Examples where joint models have succeeded

• Information extraction
  • ER+Segmentation: Poon & Domingos, AAAI07
  • SRL: Srikumar & Roth, EMNLP11
  • Within-doc extraction: Singh et al., AKBC13

• Social and communication networks
  • Fusion: Eldardiry & Neville, MLG10
  • EMailActs: Carvalho & Cohen, SIGIR05
  • GraphID: Namata et al., KDD11
GRAPH IDENTIFICATION
Transformation

Input Graph
Available but inappropriate for analysis

Output Graph
Appropriate for further analysis

Graph Identification
Motivation: Different Networks

Communication Network
Nodes: Email Address
Edges: Communication
Node Attributes: Words

Organizational Network
Nodes: Person
Edges: Manages
Node Labels: Title
Graph Identification

Input Graph: Email Communication Network

Output Graph: Social Network

Label:
- CEO
- Manager
- Assistant
- Programmer

Slides courtesy Getoor, Namata, Kok
Graph Identification

Input Graph: Email Communication Network

Output Graph: Social Network

- What’s involved?
Graph Identification

- What’s involved?
  - Entity Resolution (ER): Map input graph nodes to output graph nodes

Input Graph: Email Communication Network

Output Graph: Social Network

Slides courtesy Getoor, Namata, Kok
Graph Identification

What’s involved?
- Entity Resolution (ER): Map input graph nodes to output graph nodes
- Link Prediction (LP): Predict existence of edges in output graph
Graph Identification

- What’s involved?
  - Entity Resolution (ER): Map input graph nodes to output graph nodes
  - Link Prediction (LP): Predict existence of edges in output graph
  - Node Labeling (NL): Infer the labels of nodes in the output graph
Most work looks at these tasks in **isolation**. In graph identification, they are:

- Evidence-Dependent – Inference depend on observed input graph
e.g., ER depends on input graph
- Intra-Dependent – Inference within tasks are dependent
e.g., NL prediction depend on other NL predictions
- Inter-Dependent – Inference across tasks are dependent
e.g., LP depend on ER and NL predictions
KNOWLEDGE
GRAPH
IDENTIFICATION

Pujara, Miao, Getoor, Cohen, ISWC 2013 (best student paper)
Motivating Problem (revised)

Internet ⇒ (noisy) Extraction Graph

Large-scale IE ⇒ Knowledge Graph

 Joint Reasoning

(Pujara et al., ISWC13)
Knowledge Graph Identification

Problem:

Extraction Graph

Solution: Knowledge Graph Identification (KGI)

- Performs graph identification:
  - entity resolution
  - node labeling
  - link prediction

- Enforces ontological constraints

- Incorporates multiple uncertain sources

(Pujara et al., ISWC13)
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)

(Pujara et al., ISWC13)
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)
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)

(Annotated) Extraction Graph
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)

After Knowledge Graph Identification

(Annotated) Extraction Graph
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)
(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

\[
\begin{align*}
p \sim q &= \max(0, p + q - 1) \\
p \vee q &= \min(1, p + q) \\
\sim p &= 1 - p \\
p \Rightarrow q &= \min(1, q - p + 1)
\end{align*}
\]
Soft Logic Tutorial: Rules to Groundings

- Given a database of evidence, we can convert rule templates to instances (grounding)
- Rules are *grounded* by substituting literals into formulas

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

\[
\text{SAMEENT}(\text{Kyrgyzstan, Kyrgyz Republic}) \\
\wedge \text{LBL(\text{Kyrgyzstan, country})} \\
\Rightarrow \text{LBL(\text{Kyrgyz Republic, country})}
\]

- The soft logic interpretation assigns a “satisfaction” value to each ground rule
Soft Logic Tutorial: Groundings to Satisfaction

\[
\text{SAMEENT}(\text{Kyrgyzstan, Kyrgyz Republic}) : 0.9 \ \hat{\wedge} \\
\text{LBL}(\text{Kyrgyzstan, country}) : 0.8
\]

\[
p\hat{\lor}q = \max(0, p + q - 1)
\]

\[
\text{SAMEENT}(\text{Kyrgyzstan, Kyrgyz Republic}) \ \hat{\wedge} \\
\text{LBL}(\text{Kyrgyzstan, country}) \\
= \max(0, 0.9 + 0.8 - 1)
\]
Soft Logic Tutorial: Groundings to Satisfaction

\[ \text{SAMEENT(Kyrgyzstan, Kyrgyz Republic)} \]
\[ \tilde{\wedge} \quad \text{LBL(Kyrgyzstan, country)) : 0.7} \]
\[ \Rightarrow \quad \text{LBL(Kyrgyz Republic, country) : 0.6} \]

\[ p \tilde{\Rightarrow} q = \min(1, q - p + 1) \]

\[ \text{SAMEENT(Kyrgyzstan, Kyrgyz Republic)} \]
\[ \tilde{\wedge} \quad \text{LBL(Kyrgyzstan, country)} \]
\[ \Rightarrow \quad \text{LBL(Kyrgyz Republic, country)} \]
\[ = \min(1, 0.6 - 0.7 + 1) = 0.9 \]
Soft Logic Tutorial: Inferring Satisfaction

\[(\text{SAMEENT}(\text{Kyrgyzstan}, \text{Kyrgyz Republic}) \land \text{LBL}(\text{Kyrgyzstan, country})) : 0.7 \Rightarrow \text{LBL}(\text{Kyrgyz Republic, country}) : ?\]
Soft Logic Tutorial: Distance to Satisfaction

Distance to Satisfaction (Loss) vs. Lbl(Kyrgyz Republic, Country)
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

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

• 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 for 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

$\Rightarrow REL( E_1, E_2, R)$

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

$\Rightarrow LBL( E, L)$
PSL Rules: Entity Resolution

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

- Rules require co-referent entities to have the same labels and relations
- Creates an equivalence class of co-referent entities

**SameEnt predicate captures confidence that entities are co-referent**
PSL Rules: Ontology

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

Selectional Preference:
\[ w_O : \text{DOM}(R, L) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{LBL}(E_1, L) \]
\[ w_O : \text{RNG}(R, L) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{LBL}(E_2, L) \]

Subsumption:
\[ w_O : \text{SUB}(L, P) \quad \land \quad \text{LBL}(E, L) \quad \Rightarrow \quad \text{LBL}(E, P) \]
\[ w_O : \text{RSUB}(R, S) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{REL}(E_1, E_2, S) \]

Mutual Exclusion:
\[ w_O : \text{MUT}(L_1, L_2) \quad \land \quad \text{LBL}(E, L_1) \quad \Rightarrow \quad \neg\text{LBL}(E, L_2) \]
\[ w_O : \text{RMUT}(R, S) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \neg\text{REL}(E_1, E_2, S) \]

Adapted from Jiang et al., ICDM 2012
\[ \phi_1 \text{ CandLbl}_{\text{struct}}(\text{Kyrgyzstan, bird}) \Rightarrow \text{Lbl}(\text{Kyrgyzstan, bird}) \]

\[ \phi_2 \text{ CandRel}_{\text{pat}}(\text{Kyrgyz Rep., Asia, locatedIn}) \Rightarrow \text{Rel}(\text{Kyrgyz Rep., Asia, locatedIn}) \]

\[ \phi_3 \text{ SameEnt}(\text{Kyrgyz Rep., Kyrgyzstan}) \land \text{Lbl}(\text{Kyrgyz Rep., country}) \Rightarrow \text{Lbl}(\text{Kyrgyzstan, country}) \]

\[ \phi_4 \text{ Dom}(\text{locatedIn, country}) \land \text{Rel}(\text{Kyrgyz Rep., Asia, locatedIn}) \Rightarrow \text{Lbl}(\text{Kyrgyz Rep., country}) \]

\[ \phi_5 \text{ Mut}(\text{country, bird}) \land \text{Lbl}(\text{Kyrgyzstan, country}) \Rightarrow \neg \text{Lbl}(\text{Kyrgyzstan, bird}) \]
Probability Distribution over KGs

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

- **CandLbl** (kyrgyzstan, bird) \( \Rightarrow \) LBL(kyrgyzstan, bird)
- **Mut**(bird, country) \( \tilde{\land} \) LBL(kyrgyzstan, bird) \( \Rightarrow \neg \text{LBL(kyrgyzstan, country)} \)
- **SameEnt**(kyrgz republic, kyrgyzstan) \( \tilde{\land} \) LBL(kyrgz republic, country) \( \Rightarrow \) LBL(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>

(From Pujara et al., ISWC13)
**LinkedBrainz**

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

**MusicBrainz**

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

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

Mapping to FRBR/FOAF ontology

- **DOM**: rdfs:domain
- **RNG**: rdfs:range
- **INV**: owl:inverseOf
- **SUB**: rdfs:subClassOf
- **RSUB**: rdfs:subPropertyOf
- **MUT**: owl:disjointWith

(Pujara et al., ISWC13)
LinkedBrainz experiments

Comparisons:

- **Baseline**: Use noisy truth values as fact scores
- **PSL-EROnly**: 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-EROnly</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

(jiang, ICDM12)

Complete: Infer full knowledge graph
- Open-world model
- All possible entities, relations, labels
- Inference assigns truth value to each variable

(Pujara et al., ISWC13)
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</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>

(Pujara et al., ISWC13)
Problem: Merge domain KG to global KG
Approach: Factored Entity Resolution model

- Goal: Build a generic entity resolution model for KGs
- Build on vast amount of work on Entity Resolution
- PSL provides an easy, flexible, sophisticated models

<table>
<thead>
<tr>
<th>Local</th>
<th>Collective</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>String similarity</td>
</tr>
<tr>
<td>New Entity</td>
<td>New Entity prior</td>
</tr>
<tr>
<td>Knowledge Graph</td>
<td>Type compatibility</td>
</tr>
<tr>
<td>Domain-Specific</td>
<td>(Album length)</td>
</tr>
</tbody>
</table>
Preliminary Results

- **Task:** ER from MusicBrainz to Google KG
- **Data:**
  - 11K MusicBrainz entities (5/5-6/29/14)
  - 330K Freebase entities
  - 15.7M relations
  - 11K human labels

<table>
<thead>
<tr>
<th>Methods</th>
<th>F1</th>
<th>AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>0.734</td>
<td>0.416</td>
</tr>
<tr>
<td>+Collective</td>
<td>0.805</td>
<td>0.569</td>
</tr>
<tr>
<td>+NewEntity</td>
<td>0.840</td>
<td>0.724</td>
</tr>
</tbody>
</table>
FASTER KNOWLEDGE
GRAPH CONSTRUCTION
Partitioning
Problem: Knowledge Graphs are HUGE

(Pujara et al, AKBC13)
Solution: Partition the Knowledge Graph
Partitioning: advantages and drawbacks

• Advantages
  • Smaller problems
  • Parallel Inference
  • Speed / Quality Tradeoff

• Drawbacks
  • Partitioning large graph time-consuming
  • Key dependencies may be lost
  • New facts require re-partitioning
Key idea: Ontology-aware partitioning

- Partition the *ontology* graph, not the knowledge graph

- Induce a partitioning of the knowledge graph based on the ontology partition

(Pujara et al., AKBC13)
Considerations: Ontology-aware Partitions

• Advantages:
  • Ontology is a smaller graph
  • Ontology coupled with dependencies
  • New facts can reuse partitions

• Disadvantages:
  • Insensitive to data distribution
  • All dependencies treated equally
Refinement: include data frequency

- Annotate each ontological element with its frequency

- Partition ontology with constraint of equal vertex weights

(Pujara et al., AKBC13)
Refinement: weight edges by type

- Weight edges by their ontological importance
Experiments: Partitioning Approaches

Comparisons (6 partitions):

- **NELL**: Default promotion strategy, no KGI
- **KGI**: No partitioning, full knowledge graph model
- **Baseline**: KGI, Randomly assign extractions to partition
- **Ontology**: KGI, Edge min-cut of ontology graph
- **O+Vertex**: KGI, Weight ontology vertices by frequency
- **O+V+Edge**: KGI, Weight ontology edges by inv. frequency

<table>
<thead>
<tr>
<th></th>
<th>AUPRC</th>
<th>Running Time (min)</th>
<th>Opt. Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>NELL</td>
<td>0.765</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KGI</td>
<td>0.794</td>
<td>97</td>
<td>10.9M</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.780</td>
<td><strong>31</strong></td>
<td>3.0M</td>
</tr>
<tr>
<td>Ontology</td>
<td>0.788</td>
<td>42</td>
<td>4.2M</td>
</tr>
<tr>
<td>O+Vertex</td>
<td>0.791</td>
<td><strong>31</strong></td>
<td>3.7M</td>
</tr>
<tr>
<td>O+V+Edge</td>
<td>0.790</td>
<td><strong>31</strong></td>
<td>3.7M</td>
</tr>
</tbody>
</table>

(Pujara et al., AKBC13)
Evolving Models
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
Approximation: KGI over subset of graph
Theory: Regret of approximating update

\[ R_n(x, y_S; \hat{w}) \leq O\left( \sqrt{\frac{B\|w\|_2}{n \cdot w_p}} \|y_S - \hat{y}_S\|_1 \right) \]
Practice: Regret and Approximation Algo

```
# epochs
inference regret

Do Nothing
Random 50%
Value 50%
WLM 50%
Relational 50%
```
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
Key Collaborators