KNOWLEDGE GRAPH CONSTRUCTION

Jay Pujara

CMPS290C
4/8/2014
Talk goals!

- Problem: converting noisy text into useful knowledge

- Topics:
  - Current state-of-the-art in Information Extraction
  - Knowledge Graphs & SRL
  - PSL Models and demo
  - Tools & Datasets
Can Computers Create Knowledge?

Internet

Massive source of publicly available information

Knowledge
Computers + Knowledge = ❤️
What does it mean to create knowledge?
What do we mean by knowledge?
Defining the Questions

- Extraction
- Representation
- Reasoning and Inference
WASHINGTON (AP) — The head of the Internal Revenue Service told House Republicans on Wednesday that it would take years to provide all the documents they have subpoenaed in their probe of how the agency handled tea party groups' applications for tax-exempt status.

The comments by IRS chief John Koskinen drew a frosty response from Republicans who run the House Government Oversight and Reform Committee, one of several congressional panels investigating the controversy. The panel's chairman, Rep. Darrell Issa, R-Calif., warned him he should comply with the request "or potentially be held in contempt" of Congress, a sometimes threatened but seldom-used authority.
A Brief (Yet Helpful) Guide to Information Extraction
WASHINGTON (AP) — The head of the Internal Revenue Service told House Republicans on Wednesday that it would take years to provide all the documents they have subpoenaed in their probe of how the agency handled tea party groups' applications for tax-exempt status.

The comments by IRS chief John Koskinen drew a frosty response from Republicans who run the House Government Oversight and Reform Committee, one of several congressional panels investigating the controversy. The panel's chairman, Rep. Darrell Issa, R-Calif., warned him he should comply with the request "or potentially be held in contempt" of Congress, a sometimes threatened but seldom-used authority.
WASHINGTON (AP) — The head of the Internal Revenue Service told House Republicans on Wednesday that it would take years to provide all the documents they have subpoenaed in their probe of how the agency handled tea party groups' applications for tax-exempt status.

The comments by IRS chief John Koskinen drew a frosty response from Republicans who run the House Government Oversight and Reform Committee, one of several congressional panels investigating the controversy. The panel's chairman, Rep. Darrell Issa, R-Calif., warned him he should comply with the request "or potentially be held in contempt" of Congress, a sometimes threatened but seldom-used authority.
Understanding entities: Entity Resolution

head

Internal Revenue Service

House Republicans

Wednesday

the documents

the agency

tea party groups’

IRS chief

John Koskinen

Republicans

the House Government Oversight and Reform Committee,

congressional panels

the controversy.

The panel

chairman

Rep. Darrell Issa

him

he

the request

Congress

authority.
Understanding entities: Entity Resolution

head
IRS chief
John Koskinen
him
he

congressional panels
the controversy
the request
Congress
authority
Wednesday
the documents
the agency
Internal Revenue Service

House Republicans
they
Republicans
the House Government Oversight and Reform Committee,
The panel

chairman
Rep. Darrell Issa

tea party groups
Understanding entities: Entity Linking

head of the Internal Revenue Service
IRS chief
John Koskinen
him
he

House Republicans
they
Republicans
the House Government Oversight and Reform Committee,
The panel

chairman
Rep. Darrell Issa

Koskinen
From Wikipedia, the free encyclopedia

Koskinen is a surname originating in Finland (in Finnish, it means "small rabbit"), where it is the ninth most common surname. It may also refer to:

* Aarne Yrjö-Koskinen (1885–1951), Finnish politician, ambassador and freiherr
* Harri Koskinen (born 1970), Finnish designer
* Jari Koskinen (born 1960), Finnish politician, Minister for Agriculture and Forestry of Finland
* Johannes Koskinen (born 1954), Finnish politician (M.P., Minister of Justice)
  * Juha Koskinen, Finnish ice hockey player
  * Juha Koskinen, Finnish musician (saxist for Norther, Wintersun)
  * Juha Koskinen (footballer) (born 1972), Finnish football (soccer) player
  * Kalle Koskinen (born 1972), Finnish ice hockey player
  * Kerko Koskinen (born 1973), Finnish musician
  * Lennart Koskinen (born 1944), clergyman in the Church of Sweden, serving as Bishop in Visby
  * Mika Koskinen (born 1988), Finnish hockey player for the Sound Tigers in AHL league
  * Pasi Koskinen (born 1972), Finnish vocalist (Amorphis)
  * Petri Koskinen (born 1963), Finnish ice hockey player
  * Rolf Koskinen (born 1939), Finnish orienteering competitor, European champion
  * Sampo Koskinen (born 1979), Finnish football (soccer) player
  * Sauli Koskinen (born 1985), a Finnish TV/radio personality and entertainment reporter
  * Tapio Koskinen (born 1963), Finnish ice hockey player
  * Yrjö Sakari Yrjö-Koskinen (1839–1903), Finnish politician (senator, Finnish Party), professor, historian

Darrell Issa
From Wikipedia, the free encyclopedia

Darrell Edward Issa ( born November 1, 1953) is the Republican U.S. Representative for California's 49th congressional district, serving since 2001. The district, numbered as the 49th District during his first term, covers the northern coastal areas of San Diego County, including cities such as Oceanside, Vista, Carlsbad and Encinitas, as well as a small portion of southern Orange County.

He was formerly a CEO of Directed Electronics, a Vista, California-based manufacturer of automobile security and convenience products. The district was numbered as the 49th District during his first term and was renumbered the 49th after the 2002 Census. Since January 2011, he has served as Chairman of the House Oversight and Government Reform Committee.

As of 2015, Issa is a multi-millionaire with a net worth estimated at as much as $455 million, which, if accurate, makes him the wealthiest currently serving member of Congress.
Understanding entities: Entity Disambiguation

Koskinen
From Wikipedia, the free encyclopedia

Koskinen is a surname originating in Finland (in Finnish, it means "small rapids"), where it is the ninth most common[1] surname. It may also refer to:

- Aarno Yrjö-Koskinen (1885–1951), Finnish politician, ambassador and freiherr
- Harri Koskinen (born 1970), Finnish designer
- Jari Koskinen (born 1960), Finnish politician, Minister for Agriculture and Forestry of Finland
- Johannes Koskinen (born 1954), Finnish politician (M.P., Minister of Justice)
- Joonas Koskinen, Finnish ice hockey player
- Jukka Koskinen, Finnish musician (bassist for Norther, Wintersun)
- Jukka Koskinen (footballer) (born 1972), Finnish football (soccer) player
- Kalle Koskinen (born 1972), Finnish ice hockey player
- Kerku Koskinen (born 1973), Finnish musician
- Lennart Koskinen (born 1944), clergyman in the Church of Sweden, serving as Bishop in Visby
- Mikko Koskinen (born 1988), Finnish hockey player for the Sound Tigers in AHL league
- Pasi Koskinen (born 1972), Finnish vocalist (Amorphis)
- Petri Koskinen (born 1983), Finnish ice hockey player
- Rolf Koskinen (born 1939), Finnish orienteering competitor, European champion
- Sampo Koskinen (born 1979), Finnish football (soccer) player
- Sauli Koskinen (born 1985), a Finnish TV/radio personality and entertainment reporter
- Tapio Koskinen (born 1953), Finnish ice hockey player
- Yrjö Sakari Yrjö-Koskinen (1830–1903), Finnish politician (senator, Finnish Party), professor, historian
WASHINGTON (AP) — The head of the Internal Revenue Service told House Republicans on Wednesday that it would take years to provide all the documents they have subpoenaed in their probe of how the agency handled tea party groups' applications for tax-exempt status.

The comments by IRS chief John Koskinen drew a frosty response from Republicans who run the House Government Oversight and Reform Committee, one of several congressional panels investigating the controversy. The panel's chairman, Rep. Darrell Issa, R-Calif., warned him he should comply with the request "or potentially be held in contempt" of Congress, a sometimes threatened but seldom-used authority.
Leadership Patterns:
  _ chief _
  IRS chief John Koskinen
  _ chairman _
  The panel's chairman, Rep. Darrell Issa

Subset Patterns:
  _ one of _
  the House Government Oversight and Reform Committee, one of several congressional panels

Association Patterns:
  _, _
  Darrell Issa, R-Calif

Who is the head of the IRS?
Who chairs the House Oversight & Reform Committee?
How do the House Republicans relate to Congress?
Which state does Darrell Issa represent?
Representing knowledge from text

organizationleadbyperson(IRS, John Koskinen)

organizationleadbyperson(House Oversight & Reform Committee, Darrell Issa)

subpartoforganization(House Oversight & Reform Committee, Congress)

politicianmemberofpoliticsgroup(Darrell Issa, Republicans)

politicianholdsoffice(Darrell Issa, Representative)

locationrepresentedbypolitician(California, Darrell Issa)
Knowledge Graph representation

- Each entity is a node (red squares)
- Each node has attributes (blue circles)
- Edges between nodes represent relationships

This representation emphasizes the *relational structure* of knowledge.
Real Systems & IE Resources
NLP Toolkits

Welcome to Apache OpenNLP

The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.

It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. These tasks are usually required to build more advanced text processing services. OpenNLP also includes maximum entropy and perceptron based machine learning.

http://nlp.stanford.edu/software/

http://www.nltk.org/

http://opennlp.apache.org/

Named-entity recognition

Co-reference resolution

Parsing

Part-of-Speech Tagging

NLTK 3.0 documentation

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data.

It provides easy-to-use interfaces to over 30 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux.

Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. A new version with updates for Python 3 and NLTK 3 is in preparation.

The Stanford NLP Group makes part of our Natural Language Processing software available to everyone. These are statistical NLP toolkits for various major computational linguistics problems. They can be incorporated into applications with human language technology needs.

All the software we distribute here is written in Java. All recent distributions require Oracle Java 6+ or OpenJDK 7+. Distribution packages include components for command-line invocation, jar files, a Java API, and source code. A number of helpful people have extended our work with bindings or translations for other languages. As a result, much of this software can also easily be used from Python (or Python), Ruby, Perl, Javascript, and F# or other .NET languages.

Supported software distributions

This code is being developed, and we try to answer questions and fix bugs on a best-effort basis.

All these software distributions are open source, licensed under the GNU General Public License (v2 or later). Note that this is the full GPL, which allows many free users, but does not allow its incorporation into any type of distributed proprietary software, even in part or in translation. Commercial licensing is also available; please contact us if you are interested.

Stanford CoreNLP

An integrated suite of natural language processing tools for English and (marginally) Chinese in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. See also: Stanford Deterministic Coreference Resolution, and the online CoreNLP demo, and the CoreNLP FAQ.
Information Extraction Systems (& KBs)

**YAGO [120M]:**
Extracts primarily from structured text (Wikipedia infoboxes), with a restrictive set of relations (100) and WordNet categories

**NELL [50M]:**
Extracts from unstructured webpages (ClueWeb) with a broad set of predefined relations and categories (1000s)
[http://rtw.ml.cmu.edu/rtw/](http://rtw.ml.cmu.edu/rtw/)

**OLLIE/KnowItAll [15M/5B]:**
OpenIE - uses unstructured webpages (ClueWeb) with no predefined relations or categories
Problem Solved?
Each document is a “world” of information

- Many approaches are successful at resolving entities, and discovering relationships at the scope of a document
But what about the universe?

• Many approaches are successful at resolving entities, and discovering relationships at the scope of a document

• Building a knowledge base requires resolving entities and relationships across millions of documents
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:

- Kyrgyzstan
- 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

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** (/kærɡɪˈstɑːn/ kur-ɡɪ-staːn;[5] Kyrgyz: Кыргызстан (IPA: [qurʁɯsˈstan]); 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

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 \textit{jointly} considering multiple extractions \textit{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
Transformation

Input Graph

Available but inappropriate for analysis

Graph Identification

Output Graph

Appropriate for further analysis

Slides courtesy Getoor, Namata, Kok
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
- nsmith@msn.com
- mtaylor@email.com
- neil@email.com
- acole@email.com
- mary@email.com

Output Graph: Social Network
- Neil Smith
- Mary Taylor
- Robert Lee
- Anne Cole
- Mary Jones

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

Input Graph: Email Communication Network
Output Graph: Social Network
Graph Identification

Input Graph: Email Communication Network

Output Graph: Social Network

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
- Knowledge Graph Identification

**Solution:** Knowledge Graph Identification (KGI)
- Performs *graph identification*:
  - entity resolution
  - node labeling
  - 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: Ontology + ER

**Uncertain Extractions:**
- 0.5: Lbl(Kyrgyzstan, bird)
- 0.7: Lbl(Kyrgyzstan, country)
- 0.9: Lbl(Kyrgyz Republic, country)
- 0.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

**Extraction Graph**

- Kyrgyzstan
- Kyrgyz Republic
- country
- bird
- Bishkek
- Rel(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

(Pujara et al., ISWC13)
Illustration of KGI

Uncertain Extractions:
- 0.5: Lbl(Kyrgyzstan, bird)
- 0.7: Lbl(Kyrgyzstan, country)
- 0.9: Lbl(Kyrgyz Republic, country)
- 0.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

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

Entity Resolution:
- SameEnt(Kyrgyz Republic, Kyrgyzstan)

After Knowledge Graph Identification

(Pujara et al., ISWC13)
Modeling Knowledge
Graph Identification
Viewing KGI as a probabilistic graphical model

(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

\[
\text{SAME_ENT}(E_1, E_2) \approx \text{LBL}(E_1, L) \Rightarrow \text{LBL}(E_2, L)
\]

- 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

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

• 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

(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)

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

Relation in Knowledge Graph

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

Predicate representing uncertain relation extraction from extractor $T$

Weight for source $T$ (labels)

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

Label in Knowledge Graph

\[ \Rightarrow \text{LBL}(E, L) \]

Predicate representing uncertain label extraction from extractor $T$
PSL Rules: Entity Resolution

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

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

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

- 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

(Pujara et al., ISWC13)
**PSL Rules: Ontology**

**Inverse:**

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

**Selectional Preference:**

\[ \text{w}_O : \text{DOM}(R, L) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{LBL}(E_1, L) \]

\[ \text{w}_O : \text{RNG}(R, L) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{LBL}(E_2, L) \]

**Subsumption:**

\[ \text{w}_O : \text{SUB}(L, P) \quad \land \quad \text{LBL}(E, L) \quad \Rightarrow \quad \text{LBL}(E, P) \]

\[ \text{w}_O : \text{RSUB}(R, S) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \text{REL}(E_1, E_2, S) \]

**Mutual Exclusion:**

\[ \text{w}_O : \text{MUT}(L_1, L_2) \quad \land \quad \text{LBL}(E, L_1) \quad \Rightarrow \quad \neg\text{LBL}(E, L_2) \]

\[ \text{w}_O : \text{RMUT}(R, S) \quad \land \quad \text{REL}(E_1, E_2, R) \quad \Rightarrow \quad \neg\text{REL}(E_1, E_2, S) \]

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>LinkedBrainz</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Community-supplied data about musical artists, labels, and creative works</td>
</tr>
<tr>
<td><strong>Noise</strong></td>
<td>Realistic synthetic noise</td>
</tr>
<tr>
<td><strong>Candidate Facts</strong></td>
<td>810K</td>
</tr>
<tr>
<td><strong>Unique Labels and Relations</strong></td>
<td>27</td>
</tr>
<tr>
<td><strong>Ontological Constraints</strong></td>
<td>49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>NELL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Real-world IE system extracting general facts from the WWW</td>
</tr>
<tr>
<td><strong>Noise</strong></td>
<td>Imperfect extractors and ambiguous web pages</td>
</tr>
<tr>
<td><strong>Candidate Facts</strong></td>
<td>1.3M</td>
</tr>
<tr>
<td><strong>Unique Labels and Relations</strong></td>
<td>456</td>
</tr>
<tr>
<td><strong>Ontological Constraints</strong></td>
<td>67.9K</td>
</tr>
</tbody>
</table>
**LinkedBrainz**

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

**Music ontology**

- 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>

(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>NELL</th>
<th>NELL’s strategy: ensure ontological consistency with existing KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSL-KGI</td>
<td>Apply full Knowledge Graph Identification model</td>
</tr>
</tbody>
</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>
Scalability
Problem: Knowledge Graphs are HUGE

(Pujara et al., AKBC13)
Solution: Partition the Knowledge Graph

(Pujara et al., AKBC13)
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

(Pujara et al., AKBC13)
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

(Pujara et al., AKBC13)
Refinement: include data frequency

- Annotate each ontological element with its frequency

- Partition ontology with constraint of equal vertex weights
Refinement: weight edges by type

- Weight edges by their ontological importance

(Pujara et al., AKBC13)
Experiments: Partitioning Approaches

Comparisons (6 partitions):

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>AUPRC</th>
<th>Running Time (min)</th>
<th>Opt. Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>NELL</td>
<td>Default promotion strategy, no KGI</td>
<td>0.765</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KGI</td>
<td>No partitioning, full knowledge graph model</td>
<td>0.794</td>
<td>97</td>
<td>10.9M</td>
</tr>
<tr>
<td>baseline</td>
<td>KGI, Randomly assign extractions to partition</td>
<td>0.780</td>
<td>31</td>
<td>3.0M</td>
</tr>
<tr>
<td>Ontology</td>
<td>KGI, Edge min-cut of ontology graph</td>
<td>0.788</td>
<td>42</td>
<td>4.2M</td>
</tr>
<tr>
<td>O+Vertex</td>
<td>KGI, Weight ontology vertices by frequency</td>
<td>0.791</td>
<td>31</td>
<td>3.7M</td>
</tr>
<tr>
<td>O+V+Edge</td>
<td>KGI, Weight ontology edges by inv. frequency</td>
<td>0.790</td>
<td>31</td>
<td>3.7M</td>
</tr>
</tbody>
</table>

(Pujara et al., AKBC13)
Richer Models
Can we add more complex rules?

- The knowledge graph can have very intricate relationships between facts:

\[ \text{CandRel}(A, T, \text{AthletePlaysForTeam}) \land \text{CandRel}(T, L, \text{TeamPlaysInLeague}) \Rightarrow \text{CandRel}(A, L, \text{AthletePlaysInLeague}) \]

Can we formalize these relationships?

See:
“Learning First-Order Horn Clauses from Web Text” Schoenmackers, Etzioni, Weld, and Davis, EMNLP10
Evolving Models
Problem: Incremental Updates to KG

How do we add new extractions to the Knowledge Graph?
Naïve Approach: Full KGI over extractions
Approximation: KGI over subset of graph
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