Rule Mining for Semantifying Wikilinks
Luis Galárraga, Danai Symeonidou, Jean-Claude Moissinac

Linked Open Data Workshop
May 11th, 2015
Keyhole, Inc

Keyhole, Inc., founded in 2001, was a pioneering software development company specializing in geospatial data visualization applications and was acquired by Google in 2004.
Wikilinks in Knowledge Bases

Keyhole, Inc

Keyhole, Inc., founded in 2001, was a pioneering software development company specializing in geospatial data visualization applications and was acquired by Google in 2004.
Keyhole, Inc

*Keyhole, Inc.*, founded in 2001, was a pioneering software development company specializing in geospatial data visualization applications and was acquired by *Google* in 2004.
Wikilinks in Knowledge Bases

- 25% of the non-literal facts in DBpedia
- Signals of connection between entities
- Some are already semantified
Wikilinks in Knowledge Bases

- 25% of the non-literal facts in DBpedia
- Signals of connection between entities
- Some are already *semantified*
Wikilinks in Knowledge Bases

- 25% of the non-literal facts in DBpedia
- Signals of connection between entities
- Some are already *semantified*
Wikilinks in Knowledge Bases

- 25% of the non-literal facts in DBpedia
- Signals of connection between entities
- Some are already semantified
Wikilinks in Knowledge Bases

- 81% of wikilinks in classes Person, Place and Organisation are *unsemantified*

- Frequent semantifications:
  - Organization \(\xrightarrow{\text{linksTo}}\) Person \(\xrightarrow{\text{currentMember}}\) Organization
  - Organization \(\xrightarrow{\text{location}}\) Location \(\xrightarrow{\text{linksTo}}\) Organization
Wikilinks in Knowledge Bases

- 81% of wikilinks in classes Person, Place and Organisation are *unsemantified*.

- Some wikilinks are *unsemantifiable*.

[Diagram showing connections between keyhole, Barack Obama, Embryonic stem cell, and Google, with linksTo relationships indicated.]
Semantifying wikilinks

- Find the “meanings” of wikilinks (link prediction)
- Learn semantification rules

```
linksTo owner
Microsoft

Company

Owner

linksTo owner
ORACLE

Sun Microsystems
```

11
Semantifying wikilinks

- Find the “meanings” of wikilinks (link prediction)
- Learn semantification rules

\[
\text{linksTo}(x, y), \ is(x, \ Company), \ is(y, \ Company) \Rightarrow \ owner(x, y)
\]
Semantifying wikilinks

Use semantification rules to propose candidate relations for unsemantified wikilinks

\[ \text{linksTo}(x, y), \text{is}(x, \text{Company}), \text{is}(x, \text{Company}) \Rightarrow \text{owner}(x, y) \]
Semantifying wikilinks

Use semantification rules to propose candidate relations for unsemantified wikilinks

\[ \text{linksTo}(x, y), \text{is}(x, \text{Company}), \text{is}(x, \text{Company}) \Rightarrow \text{owner}(x, y) \]
Experimental setup

- Training dataset (DBpedia 3.8):
  - Mapping based facts + instance information
    - 4.2M facts
    - 8M rdf:type statements
    - 1.7M entities
    - Domains: Person, Place and Organization
  - Only wikilinks of entities participating in relations
    - In order to remove some unsemantifiable wikilinks
Experimental setup

• AMIE system for rule mining
  – Horn rules of the form $B_1,\ldots, B_n \rightarrow r(x, y)$
  – Suitable for potentially incomplete KBs under the Open World Assumption
  – Scales to the size current KBs
  – Thresholds: 100 positive examples, confidence 20%

• Fire predictions of the form $r(x, y)$
  – Relation $r$ is a semantification candidate for the unsemantified wikilink $x \rightarrow y$
Experimental setup

Rank multiple candidates by confidence

\[
\text{linksTo}(x, y), \text{predecessor}(y, x), \text{is}(x, \text{Monarch}), \text{is}(y, \text{Monarch}) \Rightarrow \text{parent}(x, y) \ [0.9] \\
\text{linksTo}(x, y), \text{predecessor}(y, x), \text{is}(x, \text{Monarch}), \text{is}(y, \text{Monarch}) \Rightarrow \text{successor}(x, y) \ [0.8]
\]
Experimental setup

Rank multiple candidates by confidence

\[
\text{linksTo}(x, y), \text{predecessor}(y, x), \text{is}(x, \text{Monarch}), \text{is}(y, \text{Monarch}) \Rightarrow \text{parent}(x, y) \quad [0.9]
\]

\[
\text{linksTo}(x, y), \text{predecessor}(y, x), \text{is}(x, \text{Monarch}), \text{is}(y, \text{Monarch}) \Rightarrow \text{successor}(x, y) \quad [0.8]
\]

Louis XV of France

Luis XVI of France
Experimental setup

Rank multiple candidates by confidence

linksTo(x, y), predecessor(y, x), is(x, Monarch), is(y, Monarch) => parent(x, y) [0.9]
linksTo(x, y), predecessor(y, x), is(x, Monarch), is(y, Monarch) => successor(x, y) [0.8]

Ranking of candidates: parent, predecessor
Experimental setup

- If multiple rules predict the same candidate, aggregate the confidences

\[
\text{conf}(p) = 1 - \prod_{i=1}^{|R|} 1 - \Phi(R_i, p) \times \text{conf}(R_i)
\]

- Naive approach assumes independence of rules
- It benefits candidates predicted by multiple rules
Results

- **3.5K semantification rules**

  \[
  \text{linksTo}(x, y), \text{picture}(y, x), \text{is}(x, \text{Structure}), \text{is}(y, \text{PopPlace}) \Rightarrow \text{location}(x, y) \\
  \text{linksTo}(y, x), \text{routeStart}(y, x), \text{is}(x, \text{Road}), \text{is}(y, \text{Road}) \Rightarrow \text{routeJunction}(x, y)
  \]

- **181K semantified wikilinks**
  - Around 1.8K corroborated in DBpedia 3.9

<table>
<thead>
<tr>
<th>Entity 1</th>
<th>Entity 2</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interstate 76 (west)</td>
<td>Colorado State Highway</td>
<td>routeJunction</td>
</tr>
<tr>
<td>J. Bracken Lee</td>
<td>Herbert B. Maw</td>
<td>predecessor, parent, governor</td>
</tr>
<tr>
<td>WHQX</td>
<td>WTZE</td>
<td>sisterStation</td>
</tr>
</tbody>
</table>

- **Data available at**
  [http://luisgalarraga.de/semantifying-wikilinks/](http://luisgalarraga.de/semantifying-wikilinks/)
Results

- Precision of the approach
  - Semantification candidates evaluated manually on a random sample of 60 wikilinks
  - Error margin calculated using the Wilson interval score

<table>
<thead>
<tr>
<th>Precision@1</th>
<th>Precision@3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.77 ± 0.10</td>
<td>0.67 ± 0.07</td>
</tr>
</tbody>
</table>
Conclusions & Outlook

- Wikilinks are signals of semantic connection
  - Stored in KBs but rarely used
  - Learning the semantics of such signals is a link prediction task
- Rule Mining and naive inference are a reasonable alternative to semantify wikilinks
  - Exploit the already semantified wikilinks
- Outlook
  - Extend to other KBs
  - Apply more rigorous inference approaches