TELECOM ParisTech

Institut Mines-Télécom

Rule Mining for Semantifying Wikilinks

Luis Galárraga, Danai Symeonidou, Jean-Claude Moissinac

Linked Open Data Workshop

May 11th, 2015

Wikilinks in Knowledge Bases

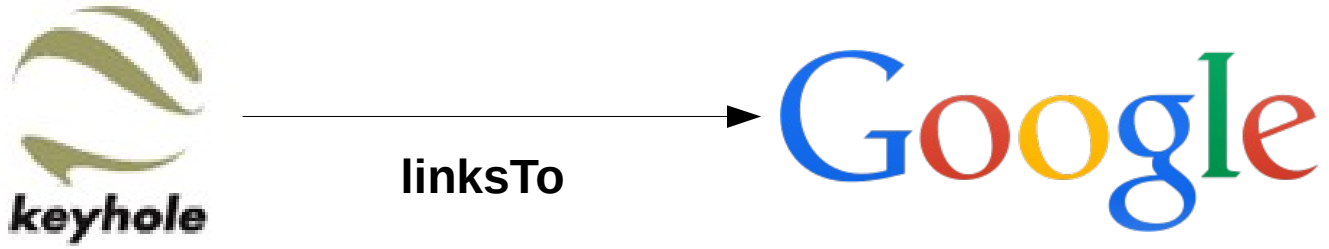
Keyhole, Inc

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

Wikilinks in Knowledge Bases

Keyhole, Inc

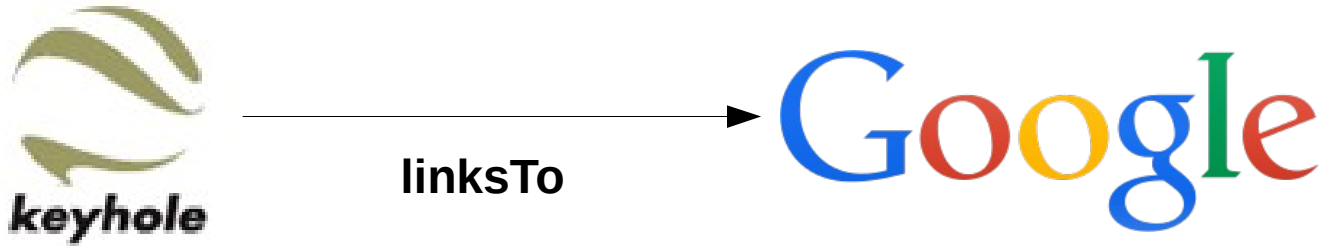
Keyhole, Inc., founded in 2001, was a pioneering software development company specializing in geospatial data visualization applications and was acquired by Google in 2004.



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.

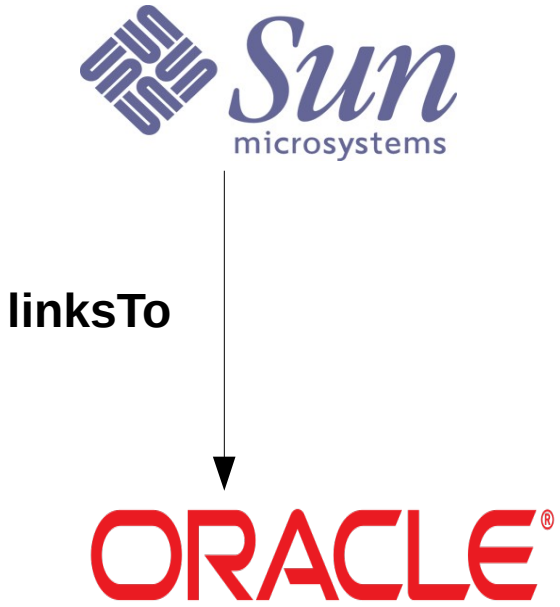


Wikilinks in Knowledge Bases

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

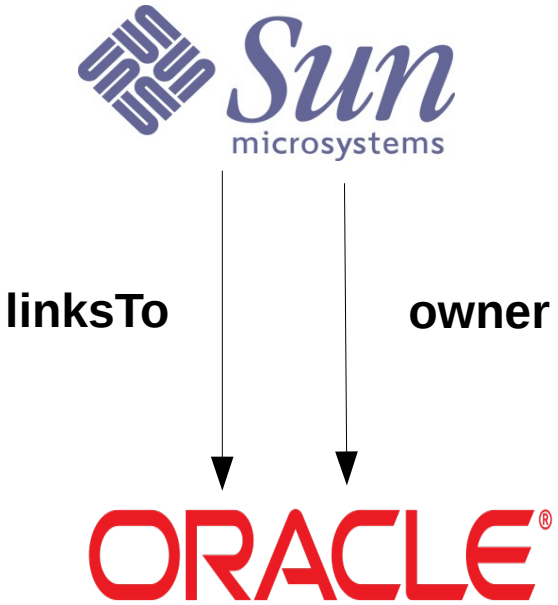
Wikilinks in Knowledge Bases

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



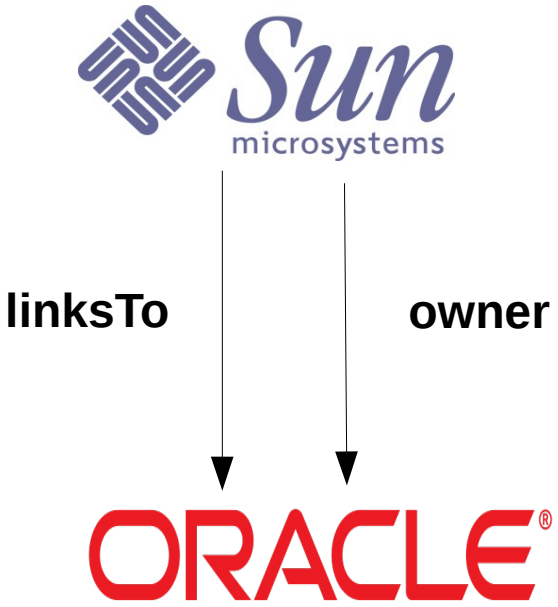
Wikilinks in Knowledge Bases

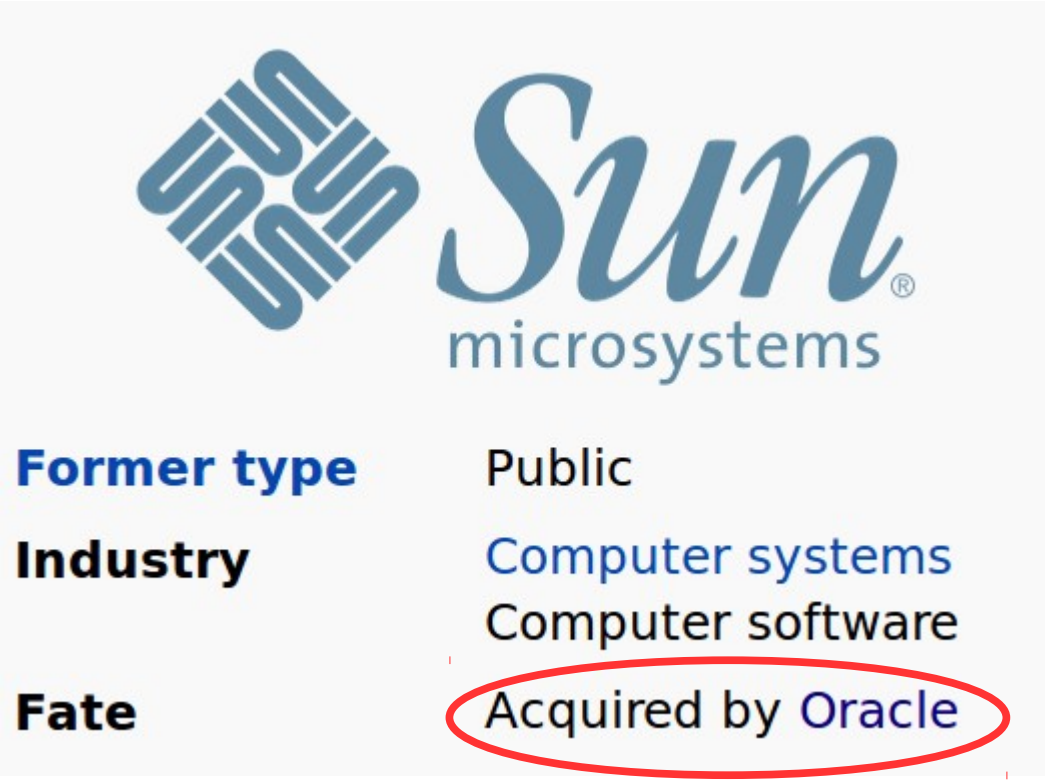
- 25% of the non-literal facts in DBpedia
- Signals of connection between entities
- Some are already *semantified*



Wikilinks in Knowledge Bases

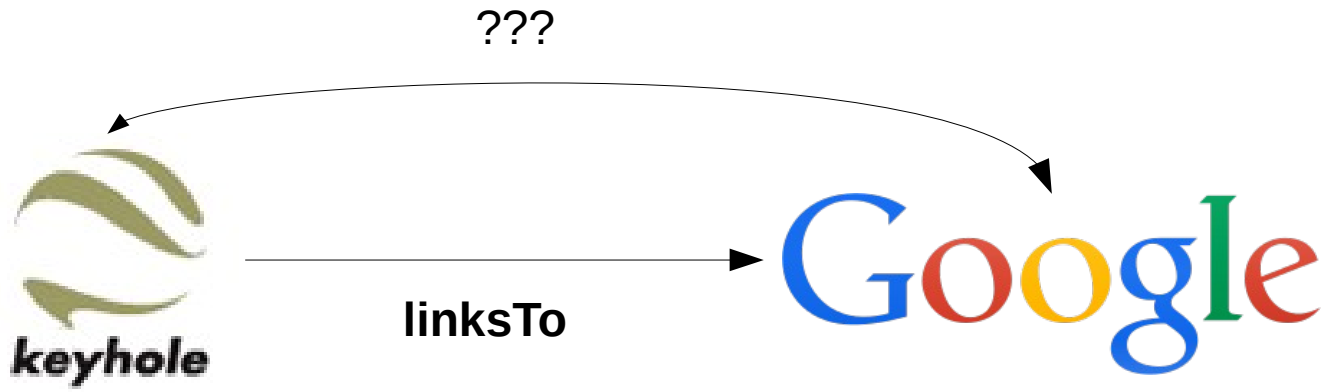
- 25% of the non-literal facts in DBpedia
- Signals of connection between entities
- Some are already *semantified*



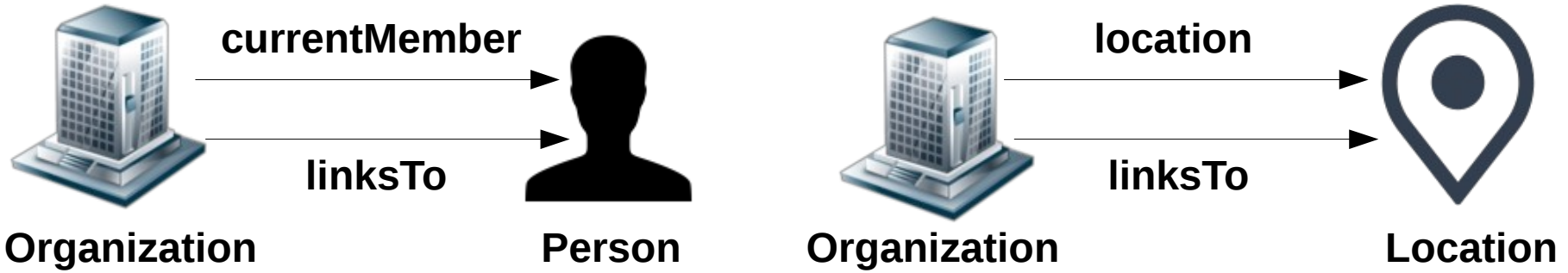


Wikilinks in Knowledge Bases

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

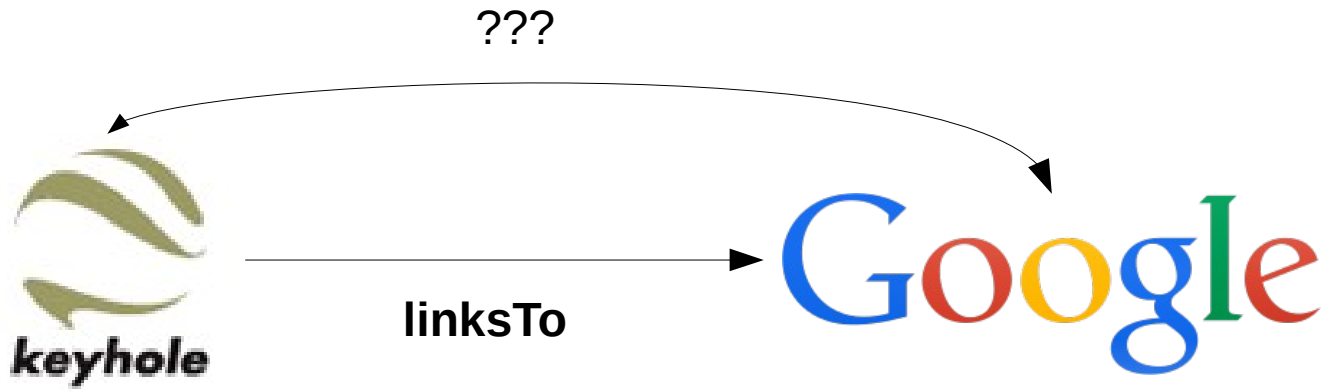


- Frequent semantifications

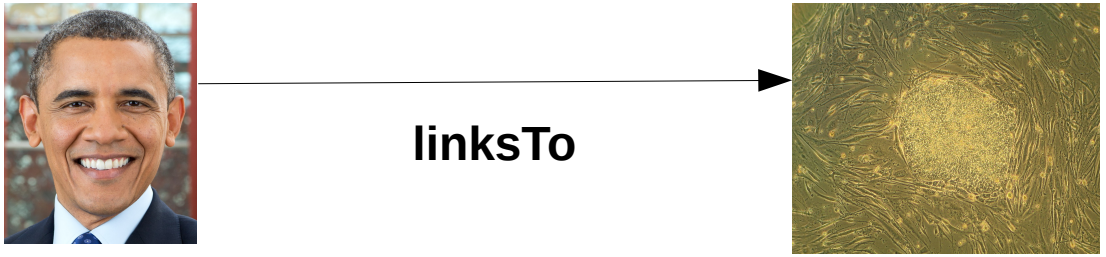


Wikilinks in Knowledge Bases

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



- Some wikilinks are *unsemantifiable*

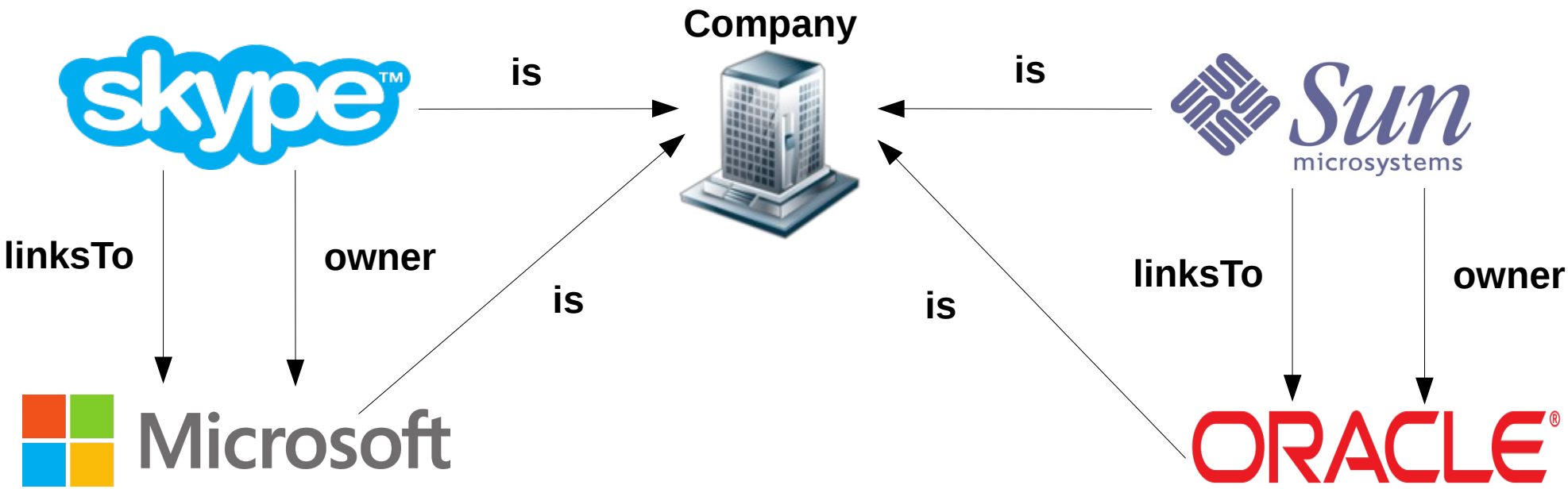


Barack Obama

Embryonic stem cell

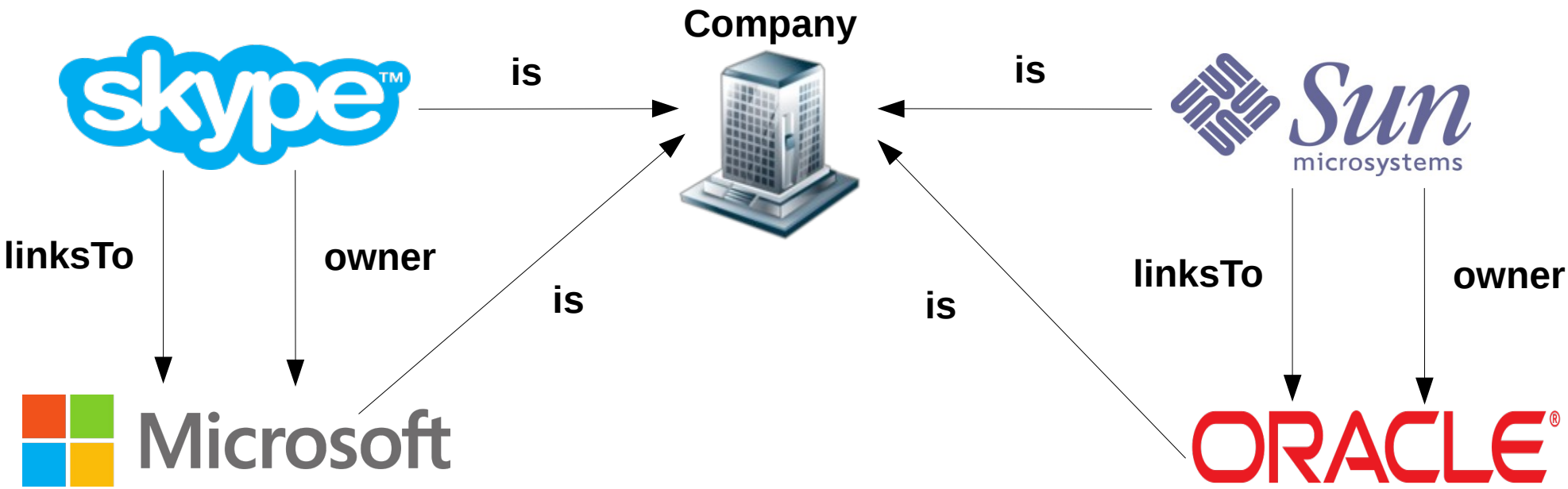
Semantifying wikilinks

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



Semantifying wikilinks

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

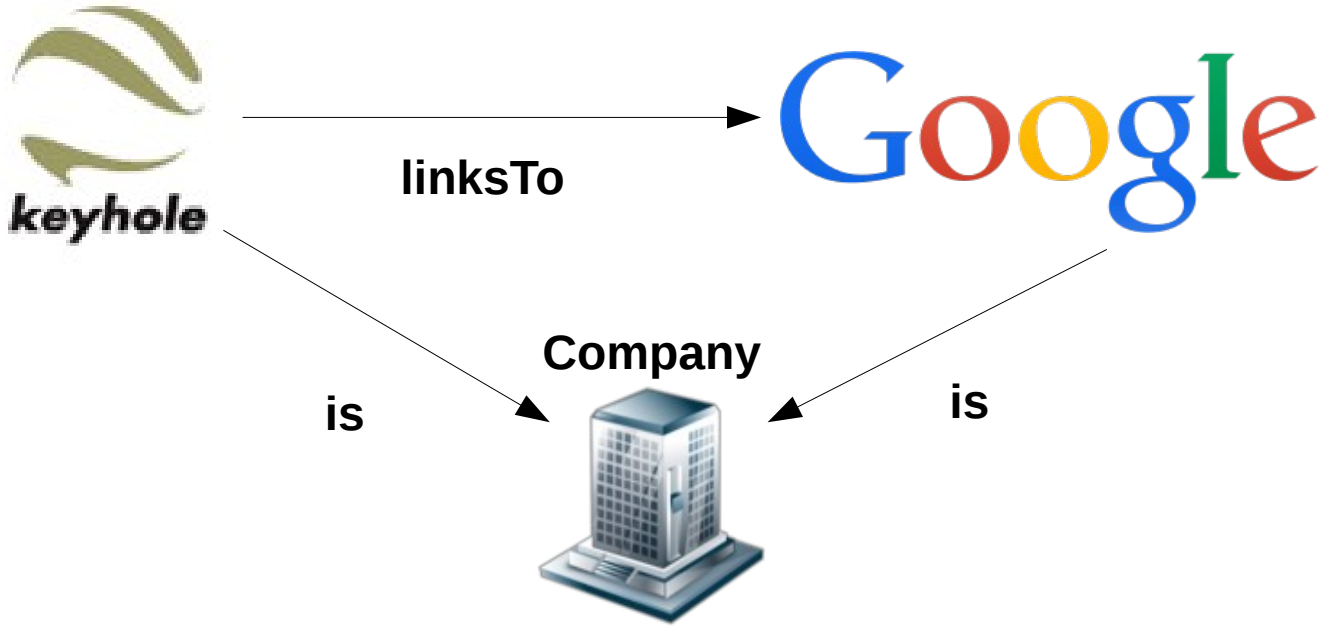


linksTo(x, y), is(x, Company), is(y, Company) => owner(x, y)

Semantifying wikilinks

Use semantification rules to propose candidate relations for unsemantified wikilinks

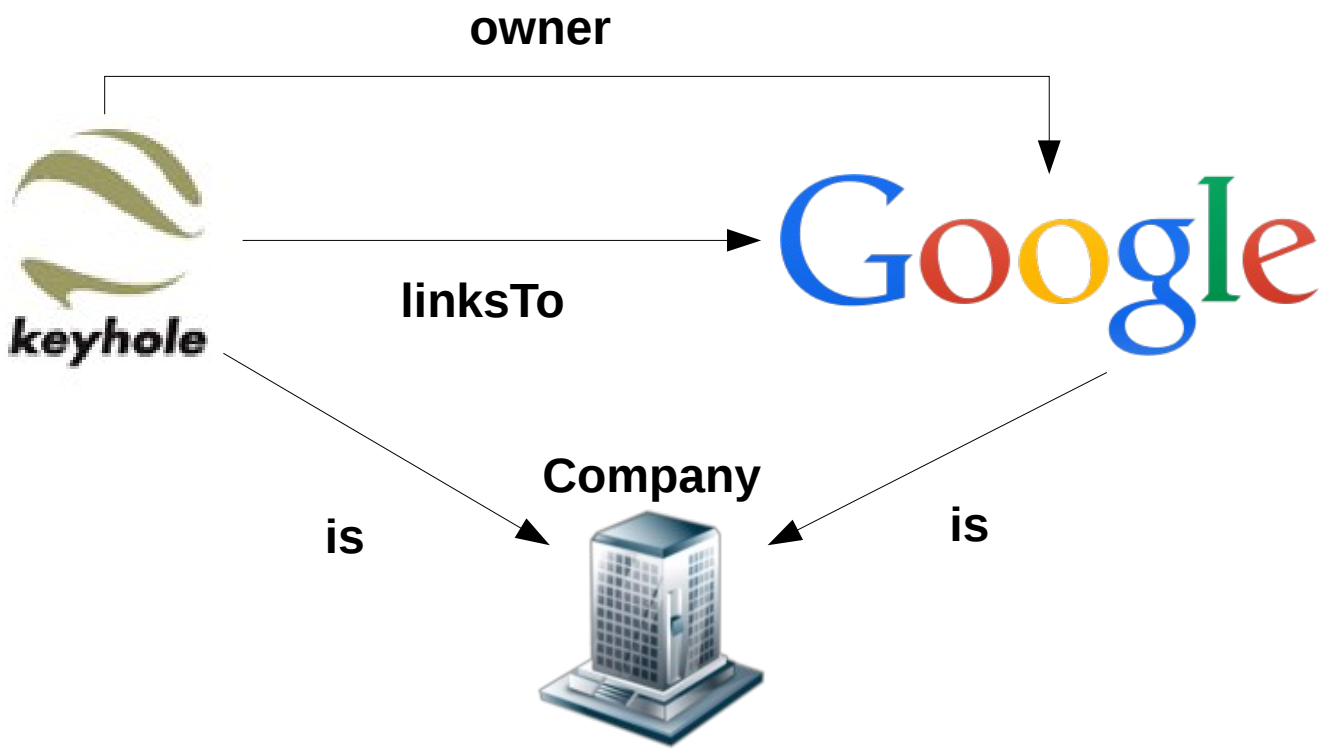
linksTo(x, y), is(x, Company), is(x, Company) => owner(x, y)



Semantifying wikilinks

Use semantification rules to propose candidate relations for unsemantified wikilinks

linksTo(x, y), is(x, Company), is(x, Company) => 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 → y

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]

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]

linksTo

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]

linksTo

parent

successor

Louis XV of France

Luis XVI of France

Ranking of candidates: parent, predecessor

Experimental setup

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

$$conf(p) = 1 - \prod_{i=1}^{|R|} 1 - \Phi(R_i, p) \times \mathrm{conf}(R_i)$$

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

Results

- 3.5K semantification rules

  linksTo(x, y), picture(y, x), is(x, Structure), is(y, PopPlace) => location(x, y)
  linksTo(y, x), routeStart(y, x), is(x, Road), is(y, Road) => routeJunction(x, y)

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

Entity 1	Entity 2	Candidates
Interstate 76 (west)	Colorado State Highway	routeJunction
J. Bracken Lee	Herbert B. Maw	predecessor, parent, governor
WHQX	WTZE	sisterStation

- Data available at 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

Precision@1	**Precision@3**
0.77 ± 0.10	0.67 ± 0.07

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