Instance-based Reasoning in Knowledge Graphs with Concepts of Neighbors

Sébastien Ferré
Joint work with Peggy Cellier and Hugo Ayats
Team LACODAM, IRISA, Univ. Rennes 1

Journée Decade, 28 juin 2022
PFIA 2022, St Étienne

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTEMES ALÉATOIRES
UMR IRISA

Scope of the Talk

- Knowledge Graphs (KG)
  - RDF graphs, possibly extended with n-ary relations
  - light RDFS-like ontologies (or none)
- "Instance-based Reasoning"
  - local forms of knowledge discovery and machine learning
    reasoning about entities by comparing them with other entities
  - like k-NN classification but with graph patterns as distances
    the same graph patterns as in SPARQL queries
- Compared to querying (SPARQL) and logical inference (RDFS)
  - same representations: graph patterns, queries, and rules
  - uncertain/statistical reasoning vs certain/logical reasoning

Overview

Instance-based reasoning questions considered in this talk:

- Notations
  - entities e, classes c, relations r, queries $Q = [x \leftarrow P]$
  - $Q(e)$ for "entity e is an answer of query Q"
  - $?x$ for unknowns
- $?Q(e_1) \wedge ?Q(e_2)$: generalization
  - *What do e_1 and e_2 have in common, as a query Q?*
- $?Q(e_1) \wedge ?Q(?e_2)$: similarity search
  - *Which entities e_2 are similar to e_1, with Q as similarity measure?*
- $r(e_1, ?e_2)$ or $r(?e_2, e_1)$: link prediction
  - *Which entity e_2 is linked to e_1 via relation r?*
  - a special case of KG completion
    - $?c(e_1)$: *What is the class of e_1?*
    - $?r(e_1, ?e_2)$, $?r(?e_2, e_1)$: *What relations connect to e_1?*
    - $?r(e_1, e_2)$: *What relation relates e_1 to e_2?*

Existing Work

- Recent trend on representation learning, aka. embeddings
  - best prediction accuracy on link prediction
  - lack of explanation for inferences
  - costly training, hence difficult with evolving data
- A long history of symbolic approaches
  - Concept Learning → generalization
    - graphs are generally propositionalized into paths
    - or seen as rooted trees (no cycle)
    - or collection of small graphs (e.g. molecules) [Kuznetsov 2013]
  - Relational Instance-Based Learning (RIBL) [Horváth 2001] → similarity
    - on rooted trees, not on KG-like relational data
    - numerical measure (edit distance)
  - Inductive Logic Programming (ILP) [Muggleton 1995], Rule Mining (AMIE+ [Galárraga 2013], AnyBURL [Meilicke 2019]) → inference
    - costly training, combinatorial search space, KGs are large

Instance-based Approach

- Reasoning principle:
  (1) e_1: entity of interest
  (2) $?Q(e_1)$: generalizations from e_1
  (3) $Q(?e_2)$: entities similar to e_1, matching some Q
  (4) $?c(e_2)$, $?r(e_2, ?e_3)$: known facts about similar entities
  (5) $c(e_1)$, $r(e_1, e_3)$: infered facts about e_1
- Formal Concept Analysis (FCA)
  - neat formalization of steps (2)-(3) (generalizations and similars)
  - concept = intension + extension
    - intension = what different objects have in common (2)
    - extension = which other objects match the intension (3)
  - limited to tabular data, intensions are sets of attributes
- Graph-FCA: extends FCA to KGs [ICFCA'15, Discr.App.Math.'20]
  - concept intensions are queries (Q)
  - concept extensions are sets of (tuples of) entities (e)
  - close to RCA (Relational Concept Analysis) [Rouane-Hacène 2013]

Overview

1. From Knowledge Graphs to Graph Concepts
2. Generalization: Conceptual Distance
3. Similarity Search: Concepts of Neighbours
4. Link Prediction
5. Application to Relation Extraction from Texts

Overview

Knowledge Graph

Entities + Relations/Classes + Triples

Diana **Charles**

Kate **William** **Harry**

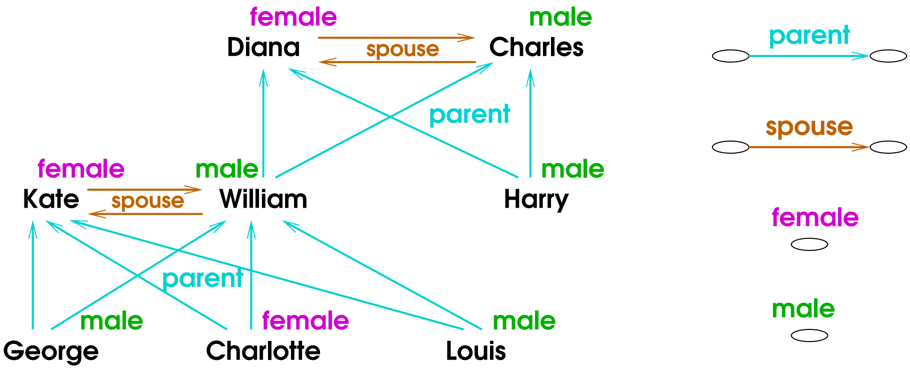
George **Charlotte** **Louis**

Knowledge Graph

Entities + Relations/Classes + Triples

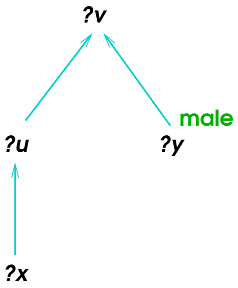
Knowledge Graph

Entities + Relations/Classes + Triples



Graph Patterns, Queries, and Answers

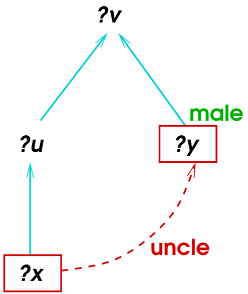
Graph Pattern



$$P = \{parent(x, u), parent(u, v), parent(y, v), male(y)\}$$

Graph Patterns, Queries, and Answers

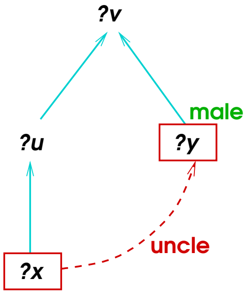
Query



$$Q = [x, y \leftarrow parent(x, u), parent(u, v), parent(y, v), male(y)]$$

Graph Patterns, Queries, and Answers

Query



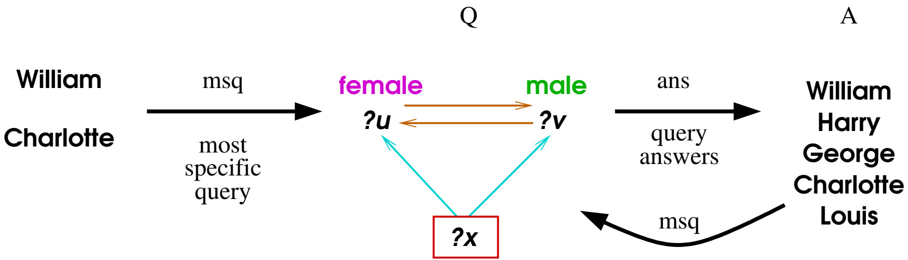
Answers (6)

x	y
George	Harry
Charlotte	Harry
Louis	Harry
George	William
Charlotte	William
Louis	William

$Q = [x, y \leftarrow parent(x, u), parent(u, v), parent(y, v), male(y)]$
$A = \{(George, Harry), (Charlotte, Harry), \ldots\}$

Graph Concepts

Starting from two entities:



A graph concept is a pair (A, Q), satisfying:

- $A = ans(Q)$: extension, set of concept instances
- $Q = msq(A)$: intension, concept description

Overview

Conceptual Distance [IC'17]

Definition

The conceptual distance between two entities e_i, e_j is defined as the most specific graph concept that contains them: $\delta(e_i, e_j) = (A_{ij}, Q_{ij})$

- $Q_{ij} = msq(\{e_i, e_j\})$: what they have in common
  this answers the question $?Q(e_i) \wedge ?Q(e_j)$
- $A_{ij} = ans(Q_{ij}) \supseteq \{e_i, e_j\}$: which entities range between them

- δ is a symbolic distance (rather than numerical)
  - distances are only partially ordered (concept inclusion)
- δ verifies distance axioms (positivity, symmetry, triangular ineq.)
  - with empty concept as zero
  - with concept union as addition
- numerical measures can be derived
  - $dist(e_i, e_j) = |ext(\delta(e_i, e_j))|$: distance as number of answers
  - $sim(e_i, e_j) = |int(\delta(e_i, e_j))|$: similarity as size of the query

Example Conceptual Distances on Mondial

On Mondial: a geographical KG with 10k entities and 12k triples

e_i - e_j	dist	$int(\delta(e_i, e_j))$ = Q = [x ← ...]
Spanish - Catalan	3	Language(x), language(Spain,x), language(Andorra, x)
Tahiti - Hawaii	6	Island(x), type(x, "volcanic"), locatedInWater(x, PacificOcean), belongsToIslands(x, ?), locatedIn(x, ?), locatedOnIsland(?, x)
Peru - Bolivia	2	Country(x), encompassed(x, America), ethnicGroup(x, Mestizo), ethnicGroup(x, European), language(x, Spanish), language(x, Quechua), language(x, Aymara), religion(x, RomanCatholic), religion(x, Protestant), wasDependentOf(x, Spain), government(x, ?), locatedIn(LakeTiticaca, x), neighbor(x, Brazil), neighbor(Brazil, x), neighbor(x, Chile), neighbor(Chile, x)

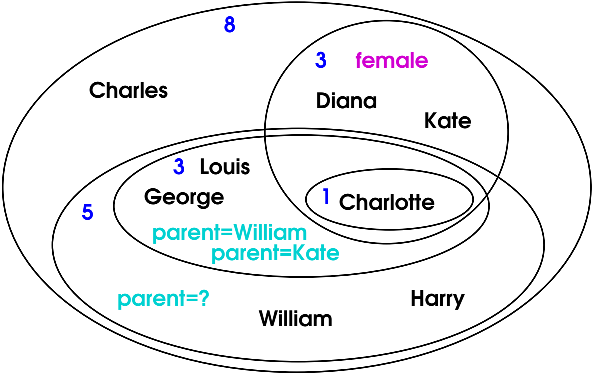
Overview

Concepts of Neighbours [IC'17]

Conceptual distances between an entity and all other entities:

$$CN(e) = \{\delta(e, e') \mid e' \in E\}$$

Example for $e = \mathit{Charlotte}$ (6 concepts of neighbors)



extensional distance

Example Similarity Searches on Mondial

Similar entities for a few seed entities, grouped by concept, in increasing extensional distance:

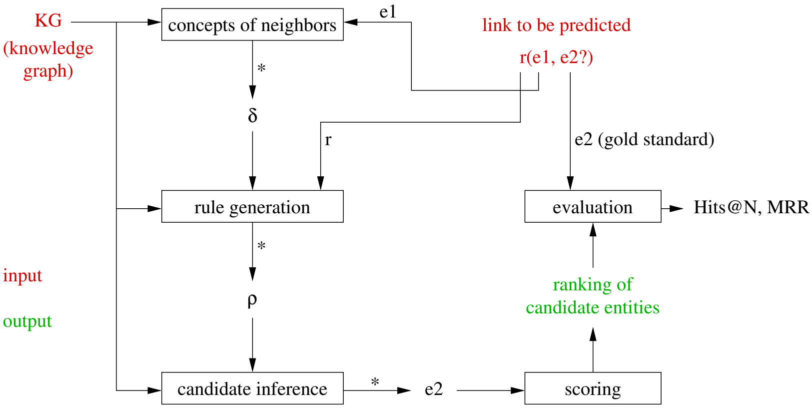
Peru	Spanish	Tahiti
Bolivia (2) *	English (2)	Mont Orohena (2)
Colombia (2) *	French (2)	Pacific Ocean (3)
Argentina (2) *	Guarani (2)	Hawaii, Hokkaido, ... (6)
Ecuador (2) *	Amerindian (2)	Bougainville, Gudalcanal, ... (12)
Panama (2) *	Miskito (2)	Taiwan, New Guinea (14)
...	Catalan, Galician (3) **	...
	Aymara, Quechua (3)	
	...	

* same extensional distance / different concepts of neighbors
⇒ different similarities/explanations

** different entities / same concept of neighbors
⇒ same similarity/explanation

Overview

Link Prediction with Concepts of Neighbors [ESWC'19]



Generation of Inference Rules from Concepts

For each $\delta = (A, Q) \in CN(e_1)$, where $Q = [x \leftarrow P]$

- Two kinds of rules are generated for the target relation r
  1. **by-copy rules**: $P \to r(x, \underline{e_2})$ for each $e_2 \in range(r)$
     - ★ x was born in Spain $\to$ x speaks <u>Spanish</u>
     - ★ inferred entities: $\{e_2\}$

$$conf := \frac{|ans([x \leftarrow P, r(x, e_2)])|}{|ans([x \leftarrow P])|}$$

  2. **by-analogy rules**: $P \to r(x, \underline{y})$ for $y \in Vars(P), y \neq x$
     - ★ x has a father z, whose wife is y $\to$ x has a mother $\underline{y}$
     - ★ inferred entities: $ans([y \leftarrow P, (x = e_1)])$

$$conf = \frac{|ans([x, y \leftarrow P, r(x, y)])|}{|ans([x, y \leftarrow P])|}$$

- Scoring+ranking: Maximum Confidence [Meilicke, 2019]

Examples of correct predictions

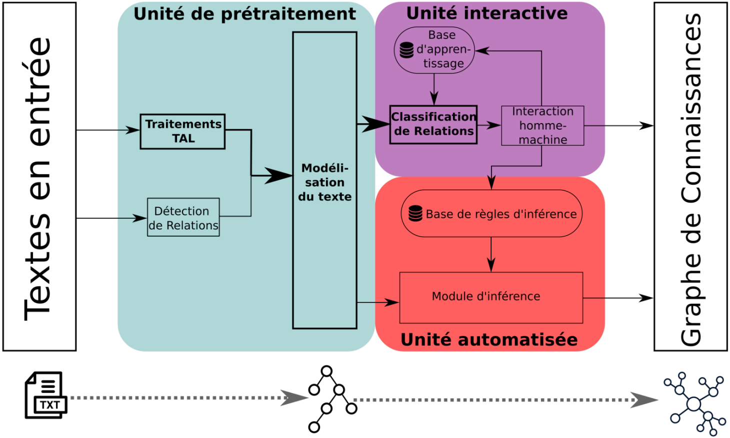
On the Mondial dataset, with timeout = 0.1+0.1s

1. mountain "Reuss" $\xrightarrow{\text{is in mountain range}}$ Alps (0.50 0.38 0.27)
   28 CNs, best explanation: *located in a place speaking Italian and German* (by-copy rule)
2. mountain "Matterhorn" $\xrightarrow{\text{is located in}}$ Switzerland (0.42 0.36 0.29)
   30 CNs, best explanation: *two mountains in the same range tend to have the same location* (by-analogy rule)

Overview

Application to Relation Extraction from Texts

[PhD work of Hugo Ayats, ICFCA'21, IDA'22]



Key role of concept intents as explanations for inferences

Conclusion and Perspectives

- Strengths
  - native interpretability of similarities and predictions (graph pattern-based rules)
  - no training phase, works on dynamic data (instance-based)
  - efficient anytime algorithm for concepts of neighbors (*progressive partitioning of the set of entities*)
- Weaknesses
  - interpretable does not mean explainable
    - post-processing concepts to extract explanations [PhD Ayats]
  - lack of approximate matching
    - on concrete domains (strings, numbers, dates)
    - on relational patterns/paths (like SPARQL path expressions)
  - efficiency on very large KGs (e.g., DBpedia)

The End

Thanks for listening !

Algorithmic and Practical Aspects

[see ESWC'18 paper on approximate query answering]

- $CNN(e, K)$ are computed by incrementally partitioning E
  - triples describing e are used as discriminating features
  - PRO: the number of clusters is bounded by $|E|$
- the partitioning algorithm is anytime
  - only coarser partition if stopped before completion
- previous experiments have shown greater efficiency compared to
  - computing conceptual distances with each entity
  - applying query relaxation to the description of e

Scoring and Ranking Inferred Entities

Maximum Confidence (introduced for AnyBURL [Meilicke, 2019])

- The score of each inferred entity e_2 is
  - the list of rule confidence measures (above 0.01)
  - in decreasing order
  - from all rules inferring e_2
  - ex: 0.94 0.86 0.33 ...
- Ranking of all inferred entities
  - in decreasing lexicographic order

e1	0.94	0.86	0.33 ...
e2	0.94	0.86	
e3	0.94	0.67	0.43 ...
e4	0.55	0.43	0.33 ...
...	...		

Experimental Results: WordNet Benchmarks

Approach	WN18 H@1	WN18 H@10	WN18 MRR	WN18RR H@1	WN18RR H@10	WN18RR MRR
Freq	1.8	5.0	2.9	1.5	4.4	2.6
Latent-based						
DISTMULT	70.1	94.3	81.3	-	-	-
ANALOGY	93.9	-	94.2	-	-	-
KB_LR	-	95.1	93.6	-	-	-
R-GCN+	69.7	96.4	81.9	-	-	-
ConvE	93.5	95.5	94.2	39.0	48.0	46.0
ComplEx-N3	-	96.0	95.0	-	**57.0**	**48.0**
CrossE	74.1	95.0	83.0	-	-	-
Rule-based						
AMIE+	87.2	94.8	-	35.8	38.8	-
RuleN	94.5	95.8	-	42.7	53.6	-
AnyBURL	93.9	95.6	95.0	**44.6**	55.5	**48.0**
C-NN (ours)	**96.7**	**97.2**	**96.9**	44.4	51.9	46.9
C-NN – best other	+2.2	+0.8	+1.9	-0.2	-5.1	-1.1
C-NN – best rule-based	+2.2	+1.4	+1.9	-0.2	-3.6	-1.1

Experimental Results: Freebase Benchmarks

	FB15k			FB15k-237		
Approach	H@1	H@10	MRR	H@1	H@10	MRR
Freq	14.3	28.5	19.2	17.5	35.6	23.6
Latent-based						
DISTMULT	52.2	81.4	63.4	10.6	37.6	19.1
ANALOGY	64.6	-	72.5	-	-	-
KB_LR	74.2	87.3	79.0	22.0	48.2	30.6
R-GCN+	60.1	84.2	69.6	15.1	41.7	24.9
ConvE	67.0	87.3	74.5	**23.9**	49.1	31.6
ComplEx-N3	-	**91.0**	**86.0**	-	**56.0**	**37.0**
CrossE	63.4	87.5	72.8	21.1	47.4	29.9
Rule-based						
AMIE+	64.7	85.8	-	17.4	40.9	-
RuleN	77.2	87.0	-	18.2	42.0	-
AnyBURL	80.4	89.0	83.0	23.0	47.9	30.0
C-NN (ours)	**82.7**	89.0	84.9	22.2	44.6	29.6
C-NN – best other	+2.3	-2.0	-1.1	-1.7	-11.4	-7.4
C-NN – best rule-based	+2.3	0.0	+1.9	-0.8	-3.3	-0.4

Relation Extraction [Ayats, Cellier, Ferré 2021]

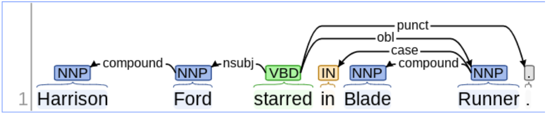
- $?r(e_1, e_2)$: Infering a relation given the subject and object entities
- Steps
  1. compute $CN(e_1, e_2)$, concepts of neighbours of the pair (e_1, e_2)
     - ★ looking for similar pairs
     - ★ graph concepts, conceptual distance, and concepts of neighbors naturally extend to tuples of entities
     - ★ concept intents are two-variable queries: $Q = [x, y \leftarrow P]$
  2. generate rules $P \rightarrow r(x, y)$ for every concept with intent $Q = [x, y \leftarrow P]$ and every relation r
  3. score and rank the relations according to rule confidences (or other scores)
- Issues:
  - there are $|E|^2$ pairs of entities to partition into concepts of neighbors
  - inference assumes the existence of a relationship from e_1 to e_2
    two pairs of unrelated entities have no reason to share something

Example of Relation Extraction

- sentence: "Harrison Ford starred in Blade Runner"
- CoreNLP processing (syntax + semantics)
- modelling as a graph: tokens as entities
- concepts of neighbors $CN(\mathrm{BladeRunner}, \mathrm{HarrisonFord})$
- Inference rules
- Inferred fact: actor(Blade Runner, Harrison Ford)

Example of Relation Extraction

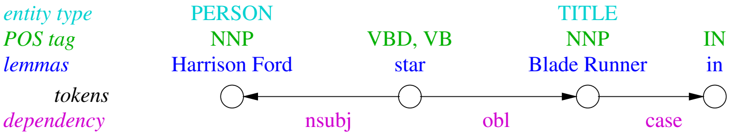
- sentence: “Harrison Ford starred in Blade Runner”
- CoreNLP processing (syntax + semantics)



- modelling as a graph: tokens as entities
- concepts of neighbors $CN(\mathrm{BladeRunner}, \mathrm{HarrisonFord})$
- Inference rules
- Inferred fact: actor(Blade Runner, Harrison Ford)

Example of Relation Extraction

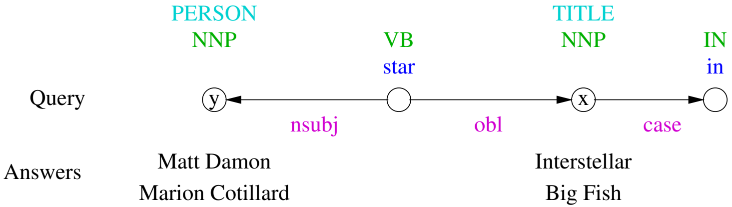
- sentence: “Harrison Ford starred in Blade Runner”
- CoreNLP processing (syntax + semantics)
- modelling as a graph: tokens as entities



- concepts of neighbors $\mathit{CN}(\mathrm{BladeRunner}, \mathrm{HarrisonFord})$
- Inference rules
- Inferred fact: actor(Blade Runner, Harrison Ford)

Example of Relation Extraction

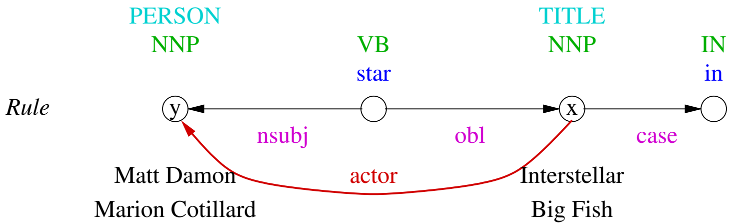
- sentence: "Harrison Ford starred in Blade Runner"
- CoreNLP processing (syntax + semantics)
- modelling as a graph: tokens as entities
- concepts of neighbors $CN(\mathrm{BladeRunner}, \mathrm{HarrisonFord})$
  only considering pairs of named entities that are related in KG



- Inference rules
- Inferred fact: actor(Blade Runner, Harrison Ford)

Example of Relation Extraction

- sentence: "Harrison Ford starred in Blade Runner"
- CoreNLP processing (syntax + semantics)
- modelling as a graph: tokens as entities
- concepts of neighbors *CN*(BladeRunner, HarrisonFord)
- Inference rules



- Inferred fact: actor(Blade Runner, Harrison Ford)

Experimental Results

- benchmark TACRED
  - collection of sentences from journalistic sources
  - each sentence has two named entities + gold standard relation
  - train/dev/test split
- relation classification:
  - accuracy = 83.6%
  - above the baseline: 80.4%
- relation extraction = detection + classification:
  - F-score = 66.9%
  - below state of the art based on BERT (F-score = 72.7%)
  - above graph convolution networks, which use similar sentence representations (F-score = 64.0-66.4%)
  - we have the benefit of explainable inferences