Towards controlled flexibility for reasoning in DLs

Anni-Yasmin Turhan

Knowledge Representation Group
Paderborn University

Joint work with: Rafael Peñaloza, Oliver Fernandez Gil

February 14th 2025

KR by Description Logics

Description Logics:

- symbolic representations by declarative formalisms
- fragments of first-order logic
- reasoning is usually decidable

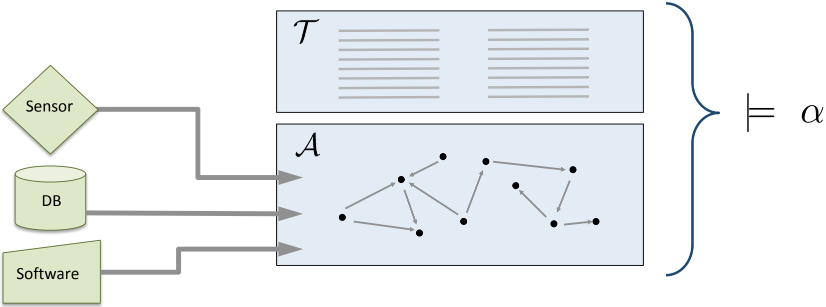
An ontology

- defines notions from an application domain
- uses words from natural language $\rightsquigarrow$ conveys meaning to humans
- relates notions to complex expressions (formulated in logic)
  $\rightsquigarrow$ gives "meaning" to notions

DL reasoning:

- well-defined
- complexity well understood
- algorithms with guarantees

Point of departure



DL knowledge bases and DL reasoning

TBox $\mathcal{T}$

$$\vdots$$

$$\text{Patient} \sqsubseteq \text{Person} \sqcap \exists \text{suffers-from}.\text{Disease}$$

$$\vdots$$

ABox $\mathcal{A}$

$\text{Person}(\text{bob})$

$\text{Disease}(\text{cooties})$

$(\text{bob}, \text{cooties})\text{suffers-from}$

Query types:

- subsumption $\quad \text{Patient} \sqsubseteq \text{Disease}$
- instance query $\quad \text{Patient}(x)$
- conjunctive query
  $\varphi(x_1, \ldots, x_n) : \exists y_1, \ldots y_m.\text{Disease}(x_1) \land \text{affects}(x_1, y_2) \ldots$

Reasoning over data

Data:

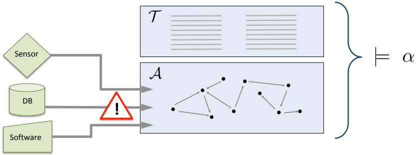
- knowledge graphs with millions of triples
- gathered automatically from different sources

:: need not be complete, accurate, consistent

How to deal with data that is

• incomplete	Use defeasible DLs
• inconsistent	Use inconsistency-tolerant semantics
• inaccurate	?

Inaccuracies?



Data and query need not fit

- Query design: precise relational structure / concept memberships not known
- exact queries may be hard to formulate
- data can evolve over time $\rightsquigarrow$ E.g. concept drift
- data sources can change over time

Relaxed queries:

- If no query answers, retrieving more than classical answers useful.
- safety-critical applications: detect instances "similar" to critical instances

How to add flexibility to reasoning to level out inaccuracies in data ?

Goal:
develop reasoning services that admit leeway for querying

Objectives:

- user is in control of relaxation
- relaxation happens local to the query (not the ontology)
- sound, complete & terminating reasoning procedures
- efficient reasoning

Classical answers – beyond and below

Relaxed Reasoning (aka: reasoning under approximate semantics):

- retrieve more than the classical answers
- retrieve also answers that are similar to classical ones
- requires: formalization of similarity

Dual notion:

Query strengthening

- retrieve some classical answers
- retrieve preferred answers
- express preferences in vocabulary of the domain

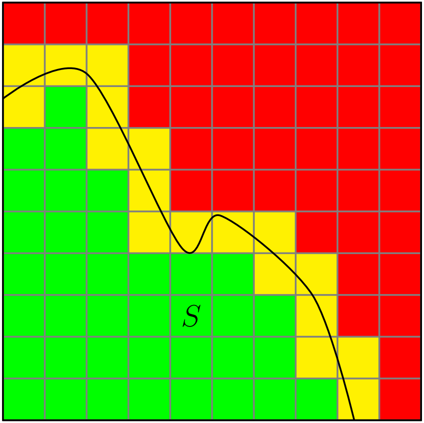
Two approaches

1. Rough DLs
   - qualitative approach
   - similarity: indiscernibility relation
2. Queries relaxed by (dis)similarity measures
   - quantitative approach
   - similarity: weighted transducer or concept similarity measure

Approach 1: Rough Description Logics

Rough DLs:

- can express qualitative form of vagueness
- semantics is based on rough sets



Description logics based on rough sets

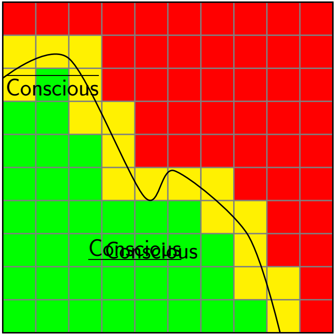
Rough description logics

- Idea: domain partitioned by indiscernibility relation ρ into granules (or: by equivalence relation $\sim$ into equivalence classes)
- populate ABox with ρ obtained from clustering algorithm
- use concept constructors:
  - lower approximation $\overline{C}$ — "strong" instances of C
  - upper approximation $\underline{C}$ — instances "similar" to those of C

Description logics based on rough sets

Rough description logics

- Idea: domain partitioned by indiscernibility relation ρ into granules (or: by equivalence relation $\sim$ into equivalence classes)
- populate ABox with ρ obtained from clustering algorithm
- use concept constructors:
  - lower approximation $\overline{C}$ — "strong" instances of C
  - upper approximation $\underline{C}$ — instances "similar" to those of C

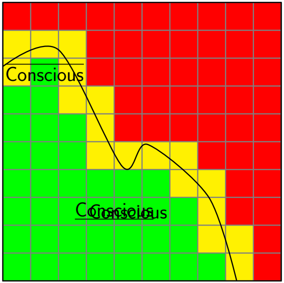


Rough $\mathcal{EL}_\bot$

$$C ::= A \mid \top \mid \bot \mid C \sqcap C \mid \exists r.C \mid \underline{C} \mid \overline{C}$$

Interpretations: $\mathcal{I} := (\Delta^\mathcal{I}, \cdot^\mathcal{I}, \sim)$

- $\Delta^\mathcal{I}$ a set; $\sim$ equivalence relation over $\Delta^\mathcal{I}$;
- $\underline{C}^\mathcal{I} := \{d \mid [d]_\sim \subseteq C^\mathcal{I}\}$;
- $\overline{C}^\mathcal{I} := \{d \mid [d]_\sim \cap C^\mathcal{I} \neq \emptyset\} = \bigcup_{\delta \in C^\mathcal{I}} [d]_\sim$

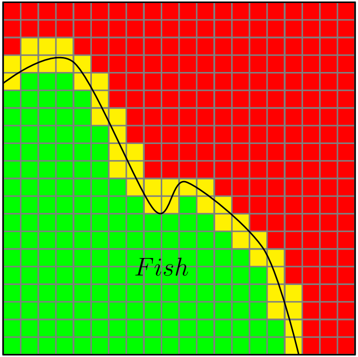


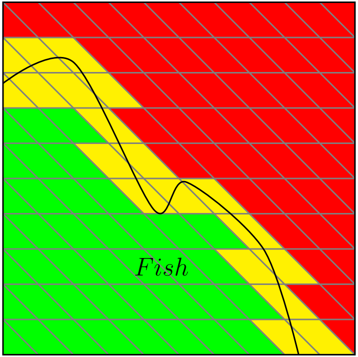
Multiple indiscernibility relations

Rough DLs are well established! $\rightsquigarrow$ Consider multiple partitions

We learn concepts through refinement ...

... or encounter different perspectives





Rough $\mathcal{EL}_\bot$ with Multiple Indiscernibility Relations

$$C ::= A \mid \top \mid \bot \mid C \sqcap C \mid \exists r.C \mid \underline{C}_i \mid \overline{C}^i$$

$$1 \leq i \leq n$$

Interpretations: $\mathcal{I} := (\Delta^\mathcal{I}, \cdot^\mathcal{I}, \sim_1, \ \dots \ , \sim_n)$

- $\underline{C}_i^\mathcal{I} := \{\delta \mid [\delta]_{\sim_i} \subseteq C^\mathcal{I}\}$;
- $\overline{C}^{i\mathcal{I}} := \{\delta \mid [\delta]_{\sim_i} \cap C^\mathcal{I} \neq \emptyset\} = \bigcup_{\delta \in C^\mathcal{I}} [\delta]_{\sim_i}$

$\rightsquigarrow$ We can refer to the different equivalence relations

Case 1: linearly ordered equivalence relations

Refining equivalence relations

- Equivalence classes just get partitioned further — not restructured!

$$\underset{\text{finest}}{\sim_1} \quad \subseteq \quad \sim_2 \quad \subseteq \quad \cdots \quad \subseteq \quad \underset{\text{coarsest}}{\sim_n}$$

- E.g. generated by one clustering algorithm and several thresholds

Reasoning in multi-rough $\mathcal{EL}_\bot$ with linearly ordered partitions

[PeñalozaT-RuleML+RR24]

- Reasoning remains in PTime!
- We constructed a completion (consequence-based) algorithm which makes consequences explicit
- Creates a canonical model

number of representatives per concept depends on n

Case 2: arbitrary set of equivalence relations

Reasoning in multi-rough $\mathcal{EL}_\perp$ is … [PeñalozaT-RuleML+RR24]

1. …in ExpTime
   (shown by reduction to $\mathcal{SHI}$(self))

   $\mathcal{SHI}$(self) can express roles with:

   - transitivity
   - symmetry (equivalence relation)
   - reflexivity

2. …ExpTime-hard
   A reduction from $\mathcal{ELI}$ ($\mathcal{EL}$ with inverse roles)

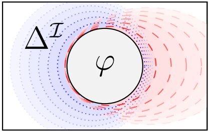
Approach 2: Relaxing queries
querying under approximate semantics

Relaxed query answer consists of:

- answer tuple and
- numerical value indicating how much tuple differs from classical answer

Advantages of relaxed queries :

- KB remains classical
- "direction" and the "degree" of relaxation is local to query
- similarity often context dependent
  $\rightsquigarrow$ intent of query?



Relaxed queries: closely related to top-k queries

How to capture similarity?

Formulation of relaxed queries:
requires (dis)similarity specification!

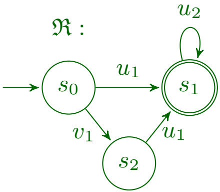
Investigated query types and approaches to model (dis)similarity:

1. Concept queries relaxed by concept similarity measures
[EckePeñalozaT- J.Appl.Logic-15]
2. Regular path queries relaxed by weighted transducers

Regular path queries

Regular path query (RPQ):

- path language specified by nondeterministic automaton (NFA) $\mathfrak{R}$ or regular expression over relations
- retrieves pairs from graph structure connected by path from $\mathcal{L}(\mathfrak{R})$
- are part of SPARQL and often used for knowledge graphs

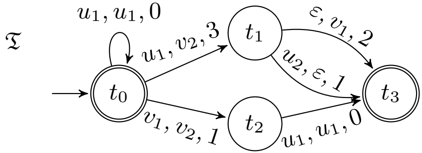


Prior work:

- answering RPQs over DL KBs by [Bienvenu-Ortiz-Simkus-JAIR-15]
- answering RPQs under approximate semantics over graph databases [Grahne-Thomo-'06]

Cost of approximating paths by weighted transducers

Weighted transducer
$\approx$ NFA with input symbol, output symbol and weight in state transitions



For approximate RPQs weighted transducer $\mathfrak{T}$ specifies:

- paths allowed as distortions of "ideal" paths required by RPQ
- corresponding distortion costs

by semi-ring $(\mathbb{N} \cup \{\infty\}, max, +, \infty, 0)$

Employing weighted transducers

Weighted transducers allow to model:

- Neutral transitions allow to preserve classical answers: $(c, d, 0)$.
- ε-transitions allow to use the edit distance as approximation:

  $(\varepsilon, s, 2) =$ delete operation $\qquad (r, \varepsilon, 1) =$ insert operation

Accepting run in $\mathfrak{T}$ on w_1 yields: output word w_2 and distortion cost

Cost of distorting w_1 into w_2 by $\mathfrak{T}$:
minimal cost from all accepting runs of $\mathfrak{T}$ that transform word u into word v.

RPQs over DL knowledge bases

Regard all models of $\mathcal{K} = (\mathcal{T}, \mathcal{A})$:
Cost of distortion in all models $\rightsquigarrow$ use supremum of costs over all models

Certain approximate answers:

$$\widetilde{\mathsf{cert}}_{\mathfrak{T}}(\mathfrak{R}(x,y),\mathcal{K}) = \Big\{(a,b,\eta) \mid \ a,b \in Ind(\mathcal{A}) \ \wedge$$
$$\eta = sup_{\mathcal{I} \models \mathcal{K}}\big\{min\{c_{\mathfrak{T}}(u,v) \mid u \in \mathcal{L}(\mathfrak{R}) \wedge a \xrightarrow{\mathcal{I},v} b\}\big\}\Big\}$$

In DLs with canonical model property:
use the universal model $\mathcal{U}_{\mathcal{K}}$ $\rightsquigarrow$ possible for $\mathcal{ELH}$, DL-Lite$_{\mathcal{R}}$

Reasoning problems for approximation semantics

τ-entailment

Instance: KB $\mathcal{K}$, transducer $\mathfrak{T}$, RPQ $\mathfrak{R}(x, y)$, tuple $\bar{a} \in \mathsf{Ind}(\mathcal{A})$ and threshold value $\mu \in \mathbb{N}$.

Question: Is $\bar{a}$ a certain approximate answer of ϕ w.r.t. $\mathcal{K}$ and $\mathfrak{T}$ with approximation cost $\eta_{\bar{a}} \leq \mu$?

cost computation

Instance: KB $\mathcal{K}$, a transducer $\mathfrak{T}$, RPQ $\mathfrak{R}(x, y)$, and a tuple $\bar{a} \in \mathsf{Ind}(\mathcal{A})$.

Output: approximation cost $\eta_{\bar{a}}$.

Complexity results

[FernandezGil-T-AAAI-20]

For $\mathcal{ELH}$ and DL-Lite$_{\mathcal{R}}$:

	τ-entailment	cost computation
2RPQs	in P	poly-time
C2RPQs	in P	poly-time

For $\mathcal{EL}$:

- bounds are tight
- hardness inherited from classical semantics [BienvenuOrtizSimkus-JAIR-15]
- approximation semantics at "no extra cost"

Conclusions

Relaxed reasoning:

- useful to adopt controlled flexibility
- qualitative and quantitative approaches investigated
- complexity is often as in classical case

Future work:

- multi-rough reasoning over data
- Which indiscernibility structure gives rise to PSpace reasoning?
- weighted transducers: other semirings?
- What happens “beyond” the Horn fragment?

Thank you!

Joint work with:
Rafael Peñaloza, Oliver Fernandez Gil