Appendix: A Framework for Exploratory Query Answering with Ontologies

Medina Andresel[1], Yazmín Ibáñez-García[2], and Magdalena Ortiz[1]

Institute of Logic and Computation, TU Wien, Austria
{andresel, ortiz}@kr.tuwien.ac.at
School of Computer Science and Informatics, Cardiff University, Cardiff, UK
ibanezgarciay@cardiff.ac.uk

Query Space Navigation

Lemma 1 *Let Ψ be a CQ template, $\mathcal{R}$ a set of reformulation axioms, and $\mathcal{O}$ an ontology. For each $\mathcal{A}$ such that $\mathcal{R}$ is valid for $(\mathcal{O}, \mathcal{A})$, $\langle Q_\Psi^\mathcal{R}, \preceq_\Psi^\mathcal{R} \rangle$ is exploratory for $\mathcal{A}$.*

Proof. Given any pair $(\mathcal{O}, \mathcal{A})$, where $\mathcal{O}$ is an ontology and $\mathcal{A}$ is an ABox such that $\mathcal{R}$ is valid for $(\mathcal{O}, \mathcal{A})$. Let $q_1, q_2 \in Q_\Psi^\mathcal{R}$ be such that $q_1 \preceq_\Psi^\mathcal{R} q_2$. Since $\mathcal{R}$ is a *DL-Lite*$^{\mathcal{HR}}$ KB and since the deriavation rules are a simplified version of the CQ rewriting rules w.r.t. a *DL-Lite*$^{\mathcal{HR}}$ ontolology presented in [1], from Proposition 3 in [1] we obtain that $q_1 \subseteq_\mathcal{R} q_2$. From validity, it must be that $q_1 \subseteq_{(\mathcal{O},\mathcal{A})} q_2$.

Complexity of Computing Maximal Neutral Reformulations

Observation 1 *Let $\mathcal{Q} = \langle Q, \preceq \rangle$ be an exploratory query space for $\mathcal{A}$ mediated by $\mathcal{O}$, and let $\preceq^c$ extend $\preceq$ so that $q_1 \preceq^c q_2$ whenever $q_1 \simeq q_2$. For $q \in Q$, we have:*

- $q_1 \in \mathsf{minStr}^{\mathsf{s}}_\mathcal{A}(q, \mathcal{Q})$ *iff* $q_1 \not\preceq^c q$, *and* $q_1 \preceq^c q_2$ *for some* $q_2 \in \mathsf{maxNeu}^{\mathsf{s}}_\mathcal{A}(q, \mathcal{Q})$.
- $q_1 \in \mathsf{minStr}^{\mathsf{g}}_\mathcal{A}(q, \mathcal{Q})$ *iff* $q_1 \not\preceq^c q$, *and* $q_2 \preceq^c q_1$ *for some* $q_2 \in \mathsf{maxNeu}^{\mathsf{g}}_\mathcal{A}(q, \mathcal{Q})$.

Proof. The observation holds due to transitivity of $\preceq^c$. For an arbitrary $q \in Q$, let q_1 denote any maximal neutral specialization of q in $\mathcal{Q}$. By definition, all minimal strict specialization candidates are obtained from choosing all $q' \preceq q_2$ for any neutral specialization q_2 of q. Suppose that $q_1 \simeq q_2$, then $q_1 \preceq^c q_2$ and $q_1 \succeq^c q_2$, by transitivity of $\succeq^c$ we obtain that $q_1 \succeq^c q'$. Lastly, we choose only those such that $q' \not\simeq q$. For minimal strict generalization case the proof is analogous.

Let $\mathcal{Q}$ be a query space. We identify the following decision problem: Given q and q in $\mathcal{Q}$, verify if q is a maximal neutral specialization (MNS) (or generalization (MNG)) of q in $\mathcal{Q}$.

Theorem (Complexity of MNS) *Verification of maximal neutral specialization is* coNP*-complete, even for query spaces consisting of only tree-shaped CQs with one answer variable and $\preceq$ induced by syntactic containment.*

Proof. Lower-bound. We reduce the problem of verifing if C is the most specific concept (MSC) for $a_1, \ldots, a_n$, given ABox $\mathcal{A}$. This amounts to verifying if $\mathcal{A} \models C(a_i)$, for each a_i, and for each D such that $\mathcal{A} \models D(a_i)$ we have that $\vdash C \sqsubseteq D$. This test is known to be coNP-complete for $\mathcal{EL}$ concepts [2].

Firstly, we extend $\mathcal{A}$ into $\mathcal{A}'$ by adding $A'(a_i)$ for each a_i, where A' is a fresh concept name. Secondly, let $\mathcal{C}$ be the set of all $\mathcal{EL}$ concepts B such that $\mathcal{A} \models B(a_i)$, and let Q be the set of all tree-shaped CQs with one answer variable of form $A'(x) \wedge cq(B(x))$, where $B \in \mathcal{C}$ and cq is a function that unfolds any $\mathcal{EL}$ atom into a CQ. For any $q_1, q_2 \in Q$, let $q_1 \preceq q_2$ if for some $\theta : vars(q_2) \to vars(q_1)$, $q_2\theta \subseteq q_1$ (i.e., the set of atoms of $q_2\theta$ is a subset of the set of atoms of q_1). Clearly, $\langle Q, \preceq \rangle$ is an exploratory query space for $\mathcal{A}'$. Lastly, C is the MSC for $a_1, \ldots, a_n$ if and only if $A'(x) \wedge cq(C(x))$ is a maximal neutral specialization of $A'(x)$.

Upper-bound. Let $\mathcal{A}$ be an ABox and $\langle Q, \prec \rangle$ be an exploratory query space for $\mathcal{A}$. Given $q_1, q_2 \in Q$, q_2 is not a maximal neutral specialization if either (i) $q_2 \not\preceq q_1$, or (ii) there exists $q' \preceq q_2$ such that $q' \simeq q$, or (iii) $ans(q_2, \emptyset, \mathcal{A}) \subsetneq ans(q_1, \emptyset, \mathcal{A}')$. Such procedure uses polynomially many calls to an NP oracle (that returns true if two queries have the same answers over $\mathcal{A}$ and false otherwise), therefore we obtain that verification of maximal neutral specializations is in coNP.

Correctness of Datalog Encoding

Lemma 2 *A CQ q is derivable from Ψ using $\mathcal{R}$ iff there exists a Π_Ψ-encoding of q.*

The proof involves showing that following claims hold. Let $(\Psi, \mathcal{R})$ be an arbitrary but fixed pair cosisting of a CQ template Ψ and set of reformulation axioms $\mathcal{R}$.

Claim 2.1 *For any CQ q such that q is derivable from Ψ using $\mathcal{R}$ there exists a Π_Ψ-encoding of q.*

Proof. Let k be the number of one-step derivations used to obtain $q(\boldsymbol{x})$ from Ψ using $\mathcal{R}$. We proceed with an inductive proof on k. If $k = 0$, then q is obtained by dropping special markers from Ψ. For substitution θ such that for each $\tau_i(\boldsymbol{y}) \in \Psi$ $\theta(U_i) = c_i$ where c_i is the Datalog constant defining τ_i we obtain that the unfolding of Π_Ψ for $(c_1, \ldots, c_n)$ has as body exactly $tr(q)$ since $\mathsf{refAt}(c_\tau, c, \boldsymbol{v_y}) \in \Pi_\Psi(D_{\Psi,\mathcal{R}})$. If $k = 1$, then there exists $\mathcal{R}_1 \subseteq \mathcal{R}$ such that $\Psi \rightsquigarrow_{\mathcal{R}_1} q$. Let $\tau \in \Psi$ be such that $\mathcal{R}_1$ is applicable to τ, atom $\alpha \in q$ be such that $\tau \rightsquigarrow_{\mathcal{R}_1} \alpha$ and c_α be the Datalog constant defining α. Then based on the form of $\mathcal{R}_1$, it easily follows from the translations of Ψ and $\mathcal{R}_1$, and derivation rules in Π_{rules} that $\mathsf{refAt}(c_\tau, c, v_{\boldsymbol{y}}) \in \Pi_\Psi(D_{\Psi,\mathcal{R}_1})$. Since for each other $\tau' \neq \tau$ in Ψ, $\mathsf{cq}(\tau') \in q$, it follows that there exists $(c_1, \ldots, c_n)$ such that c_i defines some atom in q and it is a Π_Ψ-encoding of q.

Let $\Psi \rightsquigarrow_{\mathcal{R}_1} \Psi_1 \rightsquigarrow_{\mathcal{R}_2} \cdots \rightsquigarrow_{\mathcal{R}_k} \Psi_k$ be a sequence of one-step derivations and suppose q is obtained by applying $\mathcal{R}_{k+1}$ to Ψ_k, where $\mathcal{R}_i \subseteq \mathcal{R}$, for $1 \leq i \leq k+1$.

Let q_k be the CQ obtained from Ψ_k by dropping special markers. By induction hypothesis, let $\boldsymbol{c_{q_k}}$ be the Π_Ψ-encoding of q. Let $trq_k(\boldsymbol{x})$ be the translation of q_k and let θ_k be the substitution applied to $\boldsymbol{U}$ such that $trq_k(\boldsymbol{x})$ is the body of the unfolding of Π_Ψ for $\boldsymbol{c_{q_k}}$. Therefore, for each $\tau_i(\boldsymbol{y}) \in \Psi$, $\mathsf{refAt}(c_\tau, U_i, v_{\boldsymbol{y}})\theta_k \in \Pi_\Psi(D_{\Psi,\mathcal{R}})$. Let $\tau_i^k \in \Psi_k$ be such that $\mathcal{R}_{k+1}$ is applicable to τ_i^k and let $\alpha \in q$ be such that $\tau_i^k \rightsquigarrow_{\mathcal{R}_{k+1}} \alpha$. Reasoning by cases on τ_i^k and $\mathcal{R}_{k+1}$, we obtain that for θ that maps U_i to the Datalog constant defining α and all other variables in $\boldsymbol{U}$ according to θ_k, we obtain that $\mathsf{refAt}(c_{\tau_i}, U_i, v_{\boldsymbol{y}})\theta \in \Pi_\Psi(D_{\Psi,\mathcal{R}})$, hence $\boldsymbol{U}\theta$ is a Π_Ψ-encoding of q.

For the other direction, we have:

Claim 2.2 *For any CQ q such that there exists a Π_Ψ-encoding we have that q is derivable from Ψ using $\mathcal{R}$.*

Proof. Let q be an arbitrary CQ and let $\boldsymbol{c_q}$ be the Π_Ψ-encoding of q. Since both q and Ψ have each non-join and non-answer variable replaced by "_", we obtain that only join and answer variables are present in q and in the unfolding of Π_Ψ for $\boldsymbol{c_q}$. For an arbitrary query atom $P(\boldsymbol{y}) \in q$, let $\tau \in \Psi$ be such that $\mathsf{qAtm}_\tau(P, \boldsymbol{y}) \in body(\mathsf{query}_\Psi(\boldsymbol{c_q}, \boldsymbol{x}))$. Since by unfolding Π_Ψ for $\boldsymbol{c_q}$, join variables are preserved and they cannot unifiy with null, since the rules in Π_{rules} encode exactly the template-based derivation rules, which prevent a join variable to be replaced by '_'. Then, we obtain that $\mathsf{refAt}(c_\tau, P, \boldsymbol{v_y}) \in \Pi_\Psi(D_{\Psi,\mathcal{R}})$. By induction on the number of steps needed to obtain $\mathsf{refAt}(c_\tau, P, \boldsymbol{v_y})$ it is easy to show that $\tau \rightsquigarrow_{\mathcal{R}} P^{\mathsf{x}}(\boldsymbol{y})$. Since $P(\boldsymbol{y})$ is arbitrarily chosen, we can conclude then $\Psi \rightsquigarrow_{\mathcal{R}} q$.

Theorem 1. *Let $\mathcal{Q} = (Q_\Psi^{\mathcal{R}}, \preceq_\Psi^{\mathcal{R}})$ and let $D_{all} = D_{\Psi,\mathcal{R}} \cup D_{\mathcal{A}}$. For $q, q' \in \mathcal{Q}$, let $\boldsymbol{c_q}$ be Π_Ψ-encoding of q and $\boldsymbol{c_{q'}}$ the Π_Ψ-encoding of q'. Then, for $\mathsf{x} \in \{\mathsf{s}, \mathsf{g}\}$:*
(a) $q' \in \mathsf{maxNeu}^{\mathsf{x}}_{\mathcal{A}}(q, \mathcal{Q})$ iff $\mathsf{max}_{\mathsf{x}}(\boldsymbol{c_q}, \boldsymbol{c_{q'}}) \in \Pi_{\Psi,\mathcal{R},ref}(D_{all})$,
(b) $q' \in \mathsf{minStr}^{\mathsf{x}}_{\mathcal{A}}(q, \mathcal{Q})$ iff $\mathsf{min}_{\mathsf{x}}(\boldsymbol{c_q}, \boldsymbol{c_{q'}}) \in \Pi_{\Psi,\mathcal{R},ref}(D_{all})$.

Proof. In the Datalog encoding, we first identify for each Π_Ψ-encoding, applications of any $\mathsf{refAx}(c, c')$ that produce droping or gaining at least an answer tuple (rules defining $\mathsf{str_s}$ and $\mathsf{str_g}$). If we have that some $\mathsf{refAx}(c, c')$ does not produce a strict reformulation, then we infer it is neutral (rules defining $\mathsf{ntr_g}$ and $\mathsf{nrt_s}$) and transitively close such operations. We infer it is a maximal neutral operation if no other neutral operation can be further applied (rules defininig $\mathsf{maxNtr_s}$ and $\mathsf{maxNtr_g}$). A maximal neutral reformulation, is obtained by iteratively applying on the query encoding a maximal neutral operation for each position i that denotes a template atom $\tau_i \in \Psi$ (rules defining $\mathsf{max_s}$ and $\mathsf{max_g}$). Lastly, from Observation 1, and the fact that we transitively close the neutral operations, we obtain that any strict reformulation is obtained by applying some strict operation on any maximal neutral reformulation (rules defining $\mathsf{min_s}$ and $\mathsf{min_g}$).

References

1. Andresel, M., Ibáñez-García, Y., Ortiz, M., Simkus, M.: Relaxing and restraining queries for OBDA. In: AAAI. pp. 2654–2661. AAAI Press (2019)

2. Jung, J., Lutz, C., Wolter, F.: Least general generalizations in description logic: Verification and existence. In: AAAI 2020, Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, Februrary 7 -12, 2020, New York, New York, US (2020)