Mary, What's Like All Cats?

Andreas Ecke[1⋆], Rafael Peñaloza[1,2⋆⋆], and Anni-Yasmin Turhan[1⋆⋆⋆]

[1] Institute for Theoretical Computer Science,
Technische Universität Dresden
[2] Center for Advancing Electronics Dresden
{ecke,penaloza,turhan}@tcs.inf.tu-dresden.de

In this extended abstract we report on results recently achieved for answering instance queries relaxed by concept similarity measures [3]. Traditionally, Description Logic (DL) reasoning systems only support crisp inference services, like subsumption and instances queries. The latter can be effectively used to perform different types of search tasks: Given an ABox describing a set of individuals, an instance query returns all those that are instance of the query concept Q, rejecting all others. However, often it is also interesting to consider those individuals that are not instances: Are they completely different to Q or how similar are they to Q? In cases where the original query does not retrieve any resulting individuals, those individuals that are 'very close' to being an instance can still be a good alternative. The instance queries that do not only return the instances but also those that *nearly* match the query concept are called *relaxed instance queries* [2]. A natural way to relax instance queries is by using concept similarity measures (CSMs). Such a measure $\sim$ is a function that assigns to each pair of concepts a similarity value between 0 and 1. Together with a fixed threshold t, the instance query can be relaxed by returning all individuals that are instance of a concept with a similarity value of at least t to the query concept w.r.t. $\sim$. One advantage of using CSMs as a parameter for this inference is that they can implement different notions of similarity, and regard certain features more important than others. This allows to relax queries with respect to certain features, but not others (compare Figure 1).

Example 1. Mary likes cats [6] and has a few of them as pets. As such, her view on the similarity between other animals and cats is highly influenced by how these animals behave as pets. A dog which lives inside the house, which likes getting stroked and begs for food is more similar to a cat than a wild lion in Africa. Or more formally put: Mary's view on similarity can be expressed by a CSM $\sim_{\text{Mary}}$, which weights features related to keeping animals as a pet higher than other features and therefore yields: $(\mathsf{Cat} \sim_{\text{Mary}} \mathsf{Dog}) > (\mathsf{Cat} \sim_{\text{Mary}} \mathsf{Lion})$.

Jane, Mary's best friend, is a biologist and her view on the similarity of animals is characterized by the anatomy and evolution of animals and thus resembles the biological taxonomy of animals. As such, Jane finds lions are more similar to cats than dogs, since both cats and lions belong to the *felidae* family,

⋆ Supported by DFG Graduiertenkolleg 1763 (QuantLA).
⋆⋆ Partially supported by DFG within the Cluster of Excellence 'cfAED'
⋆⋆⋆ Partially supported by the German Research Foundation (DFG) in the Collaborative Research Center 912 "Highly Adaptive Energy-Efficient Computing".

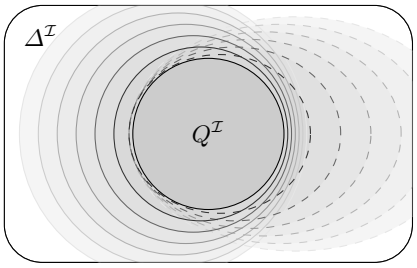

Fig. 1: Relaxed instances w.r.t. two different CSMs (solid and dashed). Darker colors represent larger thresholds t.

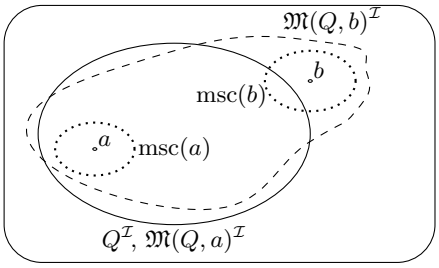

Fig. 2: Two individuals, their MSCs (dotted), and the mimics of a concept Q w.r.t. the individuals (dashed).

whereas dogs belong to the *canidae* family. Again, if we express her view on similarity as a CSM $\sim_{\text{Jane}}$, it incorporates mainly features related to the anatomy and evolution of animals and thus yields $(\mathsf{Cat} \sim_{\text{Jane}} \mathsf{Lion}) > (\mathsf{Cat} \sim_{\text{Jane}} \mathsf{Dog})$.

The CSMs $\sim_{\text{Mary}}$ and $\sim_{\text{Jane}}$ can be used to query a knowledge base describing different animals. If Mary is looking for a new pet and specifies properties of this new pet as an instance query, then using $\sim_{\text{Mary}}$ yields better results than $\sim_{\text{Jane}}$, as Mary would rather keep a dog than a lion. On the other hand, if Jane finds an animal unknown to her and wants to identify its species, a query using Jane's CSM $\sim_{\text{Jane}}$ yields better results.

We consider the Description Logic $\mathcal{EL}$ and distinguish between two kinds of TBoxes: *unfoldable TBoxes*, which are acyclic and only contain concept definitions $A \equiv C$ and *general TBoxes* which may contain GCIs $C \sqsubseteq D$. ABoxes, knowledge bases, the semantics via interpretations $\mathcal{I}$, common inferences, are defined as usual [1].

The *concept similarity measure* $\sim_{\mathcal{T}}$ w.r.t. a TBox $\mathcal{T}$ is a function of the form $\sim_{\mathcal{T}} : \mathfrak{C}(\mathcal{EL}) \times \mathfrak{C}(\mathcal{EL}) \to [0,1]$ where $\mathfrak{C}(\mathcal{EL})$ is the set of all $\mathcal{EL}$-concept descriptions and $C \sim_{\mathcal{T}} C = 1$ for all concepts C. Several properties of CSMs have been formalized in [5], the most important ones here are *symmetry* and *equivalence invariance*; the latter expresses that the similarity between two concepts does not change when replacing one concept for an equivalent one w.r.t. $\mathcal{T}$. Based on this notion we can formalize the central inference as follows:

Definition 1 (relaxed instance). *The individual a is a* relaxed instance *of the query concept Q w.r.t. the KB $\mathcal{K}$, the CSM $\sim_{\mathcal{T}}$ and the threshold $t \in [0,1)$ iff there exists a concept description X such that $Q \sim_{\mathcal{T}} X > t$ and $\mathcal{K} \models X(a)$.*

To compute the relaxed instances of an $\mathcal{EL}$-concept (w.r.t. an $\mathcal{EL}$-KB) it is not feasible to compute all sufficiently similar concepts and then perform instance checking for those, since (1) the number of those concepts can be infinite leading to an infinite number of queries and (2) a similarity measure does not necessarily provide a method how to obtain a 'sufficiently similar' concept.

The case of unfoldable TBoxes. We proceed by computing for each individual a in the ABox a concept that has the individual a as an instance and resembles C most w.r.t. $\sim_{\mathcal{T}}$. We call this the *mimic* of C w.r.t. a and $\sim_{\mathcal{T}}$, and denote it by $\mathfrak{M}(C, a)$; see Figure 2. If $\mathfrak{M}(Q, a) \sim_{\mathcal{T}} Q \geq t$ holds, then a is a relaxed instance of Q; otherwise, it cannot be a relaxed instance, as no concept can have a greater similarity value with Q while still containing a. In [2] we give a computation algorithm for mimics in $\mathcal{EL}$. The idea is to compute the role-depth bounded most specific concept k-MSC of a [7], with the role-depth of Q as the role-depth bound k, and then remove sub-concepts from the resulting concept to make it more similar to Q. This approach requires that the CSM $\sim_{\mathcal{T}}$ is symmetric, equivalence invariant, structural, i.e., it computes the similarity by induction on the structure of concepts, and is monotone in the sense that, for $N \subseteq N_C$:

$$\big(X \ \sim \prod_{A \in N} A\big) \ \geq \ \big(X \sqcap \exists r.B \ \sim \prod_{A \in N} A\big).$$

The case of general TBoxes. For general $\mathcal{EL}$-TBoxes, this role depth-based approach does not work, as the query concept may have a cyclic definition. To solve this, we introduce a CSM $\sim_c$ that uses the canonical models of a concept C w.r.t. a TBox $\mathcal{T}$, denoted with $\mathcal{I}_{C,\mathcal{T}}$, and a similarity measure $\sim_i$ between interpretations as follows: $C \sim_c D = (\mathcal{I}_{C,\mathcal{T}}, d_C) \sim_i (\mathcal{I}_{D,\mathcal{T}}, d_D)$.

The family of CSMs $\sim_c$ for $\mathcal{EL}$ inherits several formal properties from $\sim_i$, in particular symmetry and equivalence invariance. The interpretation similarity measure (ISM) $\sim_i$ can be parametrized by a weighting function that assigns different weights to each concept and role name, by a primitive measure between concept and role names and by a discounting factor.

The ISM $\sim_i$ is defined as a fixed point, and can be computed using an iterative algorithm, which converges towards the similarity value. By modifying this algorithm to generalize the pointed interpretation corresponding to the individual a to take those subsets of the concept names and role-successors that yield the highest similarity value, it actually computes the similarity between the query concept Q and the mimic of Q w.r.t. a. This is sufficient to check if a is a relaxed instance of Q w.r.t. the threshold t. This way, we get an iterative algorithm that computes all relaxed instances of Q w.r.t. $\sim_c$ and t, and that is sound and complete, i.e., it only returns individuals that are definitely relaxed instances, and it will find all relaxed instances in finitely many iterations.

To conclude, we have proposed a new reasoning service that allows relaxed instance query answering for application-specific notions of similarity by the appropriate choice of a CSM $\sim_{\mathcal{T}}$ and threshold t. We investigated necessary requirements for the CSMs to be employed. We devised computation algorithms for relaxed instances in the setting with unfoldable and with general $\mathcal{EL}$-TBoxes. For the latter setting we needed to introduce a new family of CSMs that take the whole information from general TBoxes into account. The $\sim_c$ CSMs are, to the best of our knowledge, the first CSMs of this kind for general TBoxes. Based on these we gave a computation algorithm for relaxed instances w.r.t. general TBoxes. For more details see [3, 4].

References

1. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
2. A. Ecke, R. Peñaloza, and A.-Y. Turhan. Towards instance query answering for concepts relaxed by similarity measures. In L. Godo, H. Prade, and G. Qi, editors, *Workshop on Weighted Logics for AI (in conjunction with IJCAI'13)*, Beijing, China, 2013.
3. A. Ecke, R. Peñaloza, and A.-Y. Turhan. Answering instance queries relaxed by concept similarity. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning (KR'14)*, Vienna, Austria, 2014. AAAI Press. To appear.
4. A. Ecke and A.-Y. Turhan. Similarity measures for computing relaxed instances w.r.t. general $\mathcal{EL}$-TBoxes. LTCS-Report 13-12, Chair of Automata Theory, Institute of Theoretical Computer Science, Technische Universität Dresden, Dresden, Germany, 2013. See http://lat.inf.tu-dresden.de/research/reports.html.
5. K. Lehmann and A.-Y. Turhan. A framework for semantic-based similarity measures for $\mathcal{ELH}$-concepts. In L. F. del Cerro, A. Herzig, and J. Mengin, editors, *Proc. of the 13th European Conf. on Logics in A.I. (JELIA 2012)*, Lecture Notes In Artificial Intelligence, pages 307–319. Springer, 2012.
6. C. Lutz and U. Sattler. Mary likes all cats. In F. Baader and U. Sattler, editors, *Proceedings of the 2000 International Workshop in Description Logics (DL2000)*, number 33 in CEUR-WS, pages 213–226, Aachen, Germany, August 2000. RWTH Aachen. Proceedings online available from http://SunSITE.Informatik.RWTH-Aachen.DE/Publications/CEUR-WS/Vol-33/.
7. R. Peñaloza and A.-Y. Turhan. A practical approach for computing generalization inferences in $\mathcal{EL}$. In M. Grobelnik and E. Simperl, editors, *Proceedings of the 8th European Semantic Web Conference (ESWC'11)*, Lecture Notes in Computer Science, pages 410–423. Springer, 2011.