Answering Instance Queries Relaxed by Concept Similarity

Andreas Ecke
Theoretical Computer Science,
TU Dresden, Germany

Rafael Peñaloza
Theoretical Computer Science,
TU Dresden, Germany
Center for Advancing Electronics Dresden

Anni-Yasmin Turhan
Theoretical Computer Science,
TU Dresden, Germany

{ecke,penaloza,turhan}@tcs.inf.tu-dresden.de

Abstract

In Description Logic (DL) knowledge bases (KBs) information is typically captured by crisp concepts. For many applications, querying the KB by crisp query concepts is too restrictive. A controlled way of gradually relaxing a query concept can be achieved by the use of concept similarity measures. In this paper we formalize the task of instance query answering for crisp DL KBs using concepts relaxed by concept similarity measures. We investigate computation algorithms for this task in the DL EL, their complexity and properties for the employed similarity measure regarding whether unfoldable or general TBoxes are used.

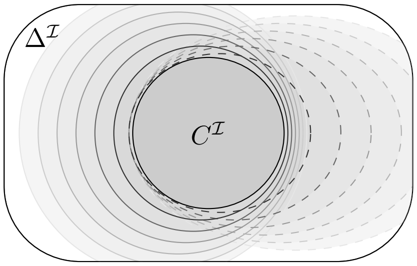

Figure 1: Relaxed instances w.r.t. two different similarity measures. Darker colors represent the relaxed instances of C w.r.t. higher degrees t.

1 Introduction

Description Logic (DL) knowledge bases are formal vocabularies that describe categories or specific subjects from application domains—such as for service matching, the biomedical or geo-spatial field. The concepts in the knowledge base are characterized by relationships to other concepts using constructors available in the DL in which the knowledge base is formulated.

The use of DLs became increasingly popular in recent years, due to the W3C standard OWL 2 for ontology languages which are based on DLs, and the development of powerful DL reasoner systems for those languages. These systems support various reasoning services, such as subsumption and instance queries. *Subsumption* determines for a pair of concepts whether elements belonging to the first concept necessarily also belong to the second one. Given a concept, *instance queries* retrieve from a knowledge base all those individuals that belong to the concept. For some applications exact matches to an instance query are sometimes too restrictive. In service matching, OWL TBoxes are employed to describe types of services and a user request for a service specifies several conditions for the desired service by a concept. For such a query concept instance query answering is performed over the OWL knowledge base that contains the individual services. If an exact match with the provided requirements is not available, a 'feasible' alternative needs to be retrieved. More precisely, the system should retrieve those individuals from the knowledge base that satisfy the main conditions, while all other conditions can be relaxed. A natural idea on how to relax the notion of instance query answering is to simply employ fuzzy DLs and perform query answering on a fuzzy variant of the initial query concept. However, on the one hand reasoning in fuzzy DLs easily becomes undecidable (Borgwardt and Peñaloza 2012) and on the other hand depending on the user and on the request, different ways of relaxing the query concept may be appropriate.

In this paper we investigate a form of answering relaxed instance queries, where the query concept can be relaxed by using concept similarity measures. A *concept similarity measure* (CSM) is a function from pairs of concepts to the unit interval, where the value 1 indicates total similarity and the value 0 total dissimilarity of the concepts. Answering *relaxed instance queries* is to compute, given concept C, a concept similarity measure $\sim$, and a degree $t \in [0,1]$, the set of individuals that instances of concepts similar to C by a degree of at least t, if measured by $\sim$. This approach, recently proposed in (Ecke, Peñaloza, and Turhan 2013), does not need to extend the DL in which the knowledge base is written, thus the complexity for standard reasoning tasks remain the same. The choice of a CSM allows to encode the application-specific notion of similarity. Furthermore the choice of t allows for a flexible degree of similarity, as depicted in Figure 1.

For DLs there is whole range of CSMs defined (see for example (Borgida, Walsh, and Hirsh 2005; d'Amato, Fanizzi, and Esposito 2005; Janowicz and Wilkes 2009;

Lehmann and Turhan 2012)), which could be employed for this task. Computing concept similarity is an ontology service that is investigated in its own right. For instance, for the Gene Ontology (Gene Ontology Consortium 2000), which is written in the DL $\mathcal{EL}$ and is used, among others, to solve the task of finding genes that realize similar functionality (Lord et al. 2003), a proliferation of different similarity measures has been defined (Lord et al. 2003; Schlicker et al. 2006; Mistry and Pavlidis 2008; Alvarez and Yan 2011).

Since concept similarity is not a formalized notion, CSMs are often defined in an ad-hoc manner or simply tuned to test data. This makes their behavior hard to predict when applied to new ontologies. In (Lehmann and Turhan 2012) a set of properties for CSMs was described and the framework given there allows users to generate CSMs for $\mathcal{EL}$-concepts with those properties. Furthermore, users can specify which parts of the vocabulary used in their knowledge base are important in regard of assessing similarity of concepts. These measures naturally allow to select which aspect of a query concept to relax.

We devise a computation algorithm for answering relaxed instance queries for concepts defined w.r.t. unfoldable $\mathcal{EL}$ TBoxes and any CSM that meets certain requirements. The core reasoning problem in our algorithm is to compute, for an individual a and the query concept C, a concept C' that *mimics* C, i.e. a concept that is ‘sufficiently similar’ to C w.r.t. the used similarity measure $\sim$ and the degree t, while preserving a as an instance.

To the best of our knowledge there is no CSM defined in the literature that takes *all* of the information from *general* $\mathcal{EL}$-TBoxes into account. Existing CSMs for general TBoxes consider only the concept hierarchy (e.g. (d’Amato, Staab, and Fanizzi 2008; Alvarez and Yan 2011)). In this paper we devise a family of CSMs that handle all the knowledge encoded in general $\mathcal{EL}$-TBoxes. We derive such CSMs by employing the canonical models of TBoxes and applying a similarity measure for interpretations. We show that this interpretation similarity measure (ISM) has properties analogous to those from (Lehmann and Turhan 2012). We define corresponding CSMs, which also preserve these properties. These CSMs for general TBoxes are the foundation to lift the computation algorithm for answering relaxed instance queries to the case of general $\mathcal{EL}$-TBoxes. It shows that the stronger CMSs obtained for this case allow for computation of relaxed instances in polynomial time, if applied to unfoldable TBoxes.

This paper is structured as follows. The next section recalls some basic notions for DLs, in particular $\mathcal{EL}$, and CSMs. Section 3 introduces relaxed instance query answering and our approach to compute it by the auxiliary task of computing a mimic. Section 4 presents a computation algorithm for relaxed instances w.r.t. unfoldable $\mathcal{EL}$-TBoxes and gives an upper bound on the complexity. Afterwards, we introduce a family of similarity measures on pointed canonical interpretations $\sim_i$ which have well-defined formal properties. From this measure we derive a CSM $\sim_c$ with the same set of properties which can be used to compute relaxed instances w.r.t. general $\mathcal{EL}$-TBoxes. We give a computation algorithm in that section and end the paper with directions for future work. Due to space constraints the proofs can be found in the technical report (Ecke and Turhan 2013).

Table 1: Concept constructors, TBox axioms and ABox assertions for $\mathcal{EL}$.

	Syntax	Semantics
top concept	$\top$	$\top^\mathcal{I} = \Delta^\mathcal{I}$
concept name	A	$A^\mathcal{I} \subseteq \Delta^\mathcal{I}$
conjunction	$C \sqcap D$	$(C \sqcap D)^\mathcal{I} = C^\mathcal{I} \cap D^\mathcal{I}$
existential restriction	$\exists r.C$	$(\exists r.C)^\mathcal{I} = \{d \in \Delta^\mathcal{I} \mid \exists e.(d,e) \in r^\mathcal{I} \wedge e \in C^\mathcal{I}\}$
concept definition	$A \equiv C$	$A^\mathcal{I} = C^\mathcal{I}$
GCI	$C \sqsubseteq D$	$C^\mathcal{I} \subseteq D^\mathcal{I}$
concept assertion	$C(a)$	$a^\mathcal{I} \in C^\mathcal{I}$
role assertion	$r(a,b)$	$(a^\mathcal{I}, b^\mathcal{I}) \in r^\mathcal{I}$

2 Preliminaries

As we want to investigate the notion of similarity in the context of the DL $\mathcal{EL}$, we briefly introduce this logic and CSMs.

The Description Logic $\mathcal{EL}$

Starting from the fixed, countably infinite, disjoint sets of concept names N_C and role names N_R, *$\mathcal{EL}$-concepts* can be constructed by the following syntactic rule

$$C, D ::= \top \mid A \mid C \sqcap D \mid \exists r.D,$$

where A is a concept name and r is a role name. The set of all $\mathcal{EL}$-concepts is denoted by $\mathfrak{C}(\mathcal{EL})$.

The semantics of concepts is defined by means of *interpretations* $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ consisting of a non-empty *domain* $\Delta^\mathcal{I}$ and an *interpretation function* $\cdot^\mathcal{I}$ that assigns binary relations on $\Delta^\mathcal{I}$ to role names and subsets of $\Delta^\mathcal{I}$ to concept names. The interpretation function is recursively extended to $\mathcal{EL}$-concepts as shown in the upper part of Table 1. We denote the class of all interpretations as $\mathfrak{I}$. A *pointed interpretation* $p = (\mathcal{I}, d)$ consists of an interpretation $\mathcal{I}$, and an element $d \in \Delta^\mathcal{I}$. $\mathfrak{P}$ is the class of all pointed interpretations, i.e., $\mathfrak{P} = \{(\mathcal{I}, d) \mid \mathcal{I} \in \mathfrak{I}, d \in \Delta^\mathcal{I}\}$. Given a pointed interpretation $p = (\mathcal{I}, d)$, the set of all $\mathcal{EL}$-concepts that have d as an instance in $\mathcal{I}$ is $\mathfrak{C}(p) = \{C \in \mathfrak{C}(\mathcal{EL}) \mid d \in C^\mathcal{I}\}$.

The middle part of Table 1 displays two kinds of concept axioms. An *$\mathcal{EL}$-TBox* or *terminology* $\mathcal{T}$ is a set of such concept axioms. An *unfoldable TBox*, is a set of concept definitions such that each concept name occurs at most once on the left-hand side of a concept definition and there are no cyclic dependencies between defined concepts. Concept names occurring on the left-hand side of a concept definition are called *defined concepts*, the other concept names are *atomic concepts*.

Consider an additional set N_I of *individual names*, which is disjoint with N_C and N_R. An interpretation $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ additionally maps each individual name to an element of

$\Delta^{\mathcal{I}}$. An ABox $\mathcal{A}$ is a set of concept or role assertions, as displayed in the lower part of Table 1. An *$\mathcal{EL}$-knowledge base* (KB) is a pair $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ where $\mathcal{T}$ is an *$\mathcal{EL}$-TBox* and $\mathcal{A}$ an *$\mathcal{EL}$-ABox*.

The semantics of interpretations is extended to TBox axioms and ABox assertions as shown in Table 1. An interpretation $\mathcal{I}$ is a model of a TBox $\mathcal{T}$ (ABox $\mathcal{A}$), if it satisfies all axioms in $\mathcal{T}$ (assertions in $\mathcal{A}$). $\mathcal{I}$ is a model of a knowledge base $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ if it is a model for both $\mathcal{T}$ and $\mathcal{A}$.

The following commonly used reasoning tasks are implemented in most DL reasoning systems. *Concept subsumption* asks, given a TBox $\mathcal{T}$ and two concepts C and D, whether C is subsumed by D w.r.t. $\mathcal{T}$ (denoted $C \sqsubseteq_{\mathcal{T}} D$), i.e. $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ for all models $\mathcal{I}$ of $\mathcal{T}$. *Concept equivalence* holds for two concepts C and D w.r.t. $\mathcal{T}$ (denoted $C \equiv_{\mathcal{T}} D$), iff $C \sqsubseteq_{\mathcal{T}} D$ and $D \sqsubseteq_{\mathcal{T}} C$. Given an individual a, a concept C, and a KB $\mathcal{K}$, a is called an *instance of* C w.r.t. $\mathcal{K}$, denoted $\mathcal{K} \models C(a)$, iff $a^{\mathcal{I}} \in C^{\mathcal{I}}$ for all models $\mathcal{I}$ of $\mathcal{K}$. Given a KB $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ and a concept C, *instance retrieval* returns all individuals from $\mathcal{A}$ that are instances of C. These reasoning tasks can all be characterized by means of simulations between interpretations.

Definition 1 (simulation)**.** Let $\mathcal{I}$ and $\mathcal{J}$ be interpretations. A relation $S \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{J}}$ is a *simulation* between $\mathcal{I}$ and $\mathcal{J}$, if the following two conditions hold:

1. For all $(d, e) \in S$ and $A \in N_C$, if $d \in A^{\mathcal{I}}$ then $e \in A^{\mathcal{J}}$.
2. For all $(d, e) \in S$, $r \in N_R$ and $(d, d') \in r^{\mathcal{I}}$, there is an $(e, e') \in r^{\mathcal{J}}$ with $(d', e') \in S$.

The pointed interpretation $p = (\mathcal{I}, d)$ *simulates* the pointed interpretation $q = (\mathcal{J}, e)$ ($p \lesssim q$), if there exists a simulation $S \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{J}}$ between $\mathcal{I}$ and $\mathcal{J}$ with $(d, e) \in S$. If $p \lesssim q$ and $q \lesssim p$, then p and q are *equisimilar* ($p \simeq q$). There is a strong connection between simulations between pointed interpretations and their concept sets.

Theorem 2 ((Lutz and Wolter 2010))**.** *Let $p, q \in \mathfrak{P}$. Then: 1. $p \lesssim q$ iff $\mathfrak{C}(p) \subseteq \mathfrak{C}(q)$, and 2. $p \simeq q$ iff $\mathfrak{C}(p) = \mathfrak{C}(q)$.*

For $\mathcal{EL}$ most reasoning procedures rely on the fact that canonical models can be built, from which entailments can be read-off directly. By $\text{Sig}(X)$ we denote the set of names and by $\mathsf{sub}(X)$ the sub-concepts appearing in X.

Definition 3. Let $C \in \mathfrak{C}(\mathcal{EL})$ and $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be an $\mathcal{EL}$-KB. The *canonical model* $\mathcal{I}_{C,\mathcal{T}} = (\Delta^{\mathcal{I}_{C,\mathcal{T}}}, \cdot^{\mathcal{I}_{C,\mathcal{T}}})$ *of* C w.r.t. the TBox $\mathcal{T}$ is defined as follows:

- $\Delta^{\mathcal{I}_{C,\mathcal{T}}} = \{d_C\} \cup \{d_D \mid \exists r.D \in \mathsf{sub}(C) \cup \mathsf{sub}(\mathcal{T})\}$
- $A^{\mathcal{I}_{C,\mathcal{T}}} = \{d_D \mid D \sqsubseteq_{\mathcal{T}} A\}$, and
- $r^{\mathcal{I}_{C,\mathcal{T}}} = \{(d_D, d_E) \mid D \sqsubseteq_{\mathcal{T}} \exists r.E\}$.

The *canonical model* $\mathcal{I}_{\mathcal{K}} = (\Delta^{\mathcal{I}_{\mathcal{K}}}, \cdot^{\mathcal{I}_{\mathcal{K}}})$ *of the KB* $\mathcal{K}$ is defined as follows:

- $\Delta^{\mathcal{I}_{\mathcal{K}}} = \{d_a \mid a \in \text{Sig}(\mathcal{A}) \cap N_I\} \ \cup$ $\{d_C \mid \exists r.C \in \mathsf{sub}(\mathcal{A}) \cup \mathsf{sub}(\mathcal{T})\}$,
- $A^{\mathcal{I}_{\mathcal{K}}} = \{d_D \mid D \sqsubseteq_{\mathcal{T}} A\} \cup \{d_a \mid \mathcal{K} \models A(a)$,
- $r^{\mathcal{I}_{\mathcal{K}}} = \{(d_D, d_E) \mid D \sqsubseteq_{\mathcal{T}} \exists r.E\} \ \cup$ $\{(d_a, d_D) \mid \mathcal{K} \models \exists r.D(a)\} \ \cup \ \{(d_a, d_b) \mid r(a, b) \in \mathcal{A}\}$.

Note that canonical models for $\mathcal{EL}$ are always finite. The canonical model $\mathcal{I}_{C,\mathcal{T}}$ can be seen as the most general model for C and $\mathcal{T}$, as any other model $\mathcal{J}$ of $\mathcal{T}$ with a $d \in C^{\mathcal{J}}$ can be simulated by $(\mathcal{I}_{C,\mathcal{T}}, d_C)$. Similarly, for any other model $\mathcal{J}$ of $\mathcal{K}$ with $d = a^{\mathcal{J}}$ for an individual a, $(\mathcal{J}, d)$ is simulated by $(\mathcal{I}_{\mathcal{K}}, d_a)$.

Theorem 4 ((Lutz and Wolter 2010))**.** *Let $\mathcal{T}$ be an $\mathcal{EL}$-TBox, C and D be $\mathcal{EL}$-concepts. Then:*

1. *for all models $\mathcal{I}$ of $\mathcal{T}$ and all elements $d \in \Delta^{\mathcal{I}}$ it is the case that $d \in C^{\mathcal{I}}$ iff $(\mathcal{I}_{C,\mathcal{T}}, d_C) \lesssim (\mathcal{I}, d)$; and*
2. *$C \sqsubseteq_{\mathcal{T}} D$ iff $d_C \in D^{\mathcal{I}_{C,\mathcal{T}}}$ (i.e., $D \in \mathfrak{C}((\mathcal{I}_{C,\mathcal{T}}, d_C))$) iff $(\mathcal{I}_{D,\mathcal{T}}, d_D) \lesssim (\mathcal{I}_{C,\mathcal{T}}, d_C)$.*

When testing whether an individual is a relaxed instance of some concept C, we will see the need to compute 'the best-fitting $\mathcal{EL}$-concept description' that has the individual as an instance. This can be realized by the task of computing the most specific concept.

Definition 5 (most specific concept)**.** Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a KB and a an individual from $\mathcal{A}$. A concept C is the *most specific concept (msc) of a w.r.t. $\mathcal{K}$* (denoted $\text{msc}_{\mathcal{K}}(a)$) if it satisfies:

1. $\mathcal{K} \models C(a)$, and
2. $\mathcal{K} \models D(a)$ implies $C \sqsubseteq_{\mathcal{T}} D$ for all concepts D.

Since the msc may not exists due to an infinite nesting of existential restrictions, one can compute an approximation by bounding the depth of the quantifiers. Formally, given a concept C, we define the role-depth $\text{rd}(C)$ of C by

$$\text{rd}(C) := \begin{cases} 0 & \text{if } C \in N_C \cup \{\top\} \\ 1 + \text{rd}(D) & \text{if } C = \exists r.D \\ max\{\text{rd}(C_1), \text{rd}(C_2)\} & \text{if } C = C_1 \sqcap C_2 \end{cases}$$

If in Definition 5 C and D have a role-depth limited to $k \in \mathbb{N}$, then C is the *role-depth bounded msc* of a w.r.t. $\mathcal{K}$ (k-$\text{msc}_{\mathcal{K}}(a)$). The msc and the k-msc are unique up to equivalence in $\mathcal{EL}$, if they exist (Peñaloza and Turhan 2011).

Concept Similarity Measures

Given a DL $\mathcal{L}$ and a TBox $\mathcal{T}$, a concept similarity measure (CSM) is a function $\sim_{\mathcal{T}} : \mathfrak{C}(\mathcal{L}) \times \mathfrak{C}(\mathcal{L}) \to [0, 1]$ such that $C \sim_{\mathcal{T}} C = 1$ for all concepts C. Intuitively, $\sim_{\mathcal{T}}$ expresses how close two concepts are. A value $C \sim_{\mathcal{T}} D = 0$ means that the concepts C and D are totally dissimilar w.r.t. $\mathcal{T}$, while a value of 1 indicates total similarity. We often simply write $\sim$ instead of $\sim_{\mathcal{T}}$ if the TBox $\mathcal{T}$ is clear from the context.

In (Lehmann and Turhan 2012) a set of properties for CSMs is defined and a framework is devised that allows to construct CSMs for $\mathcal{EL}$-concepts, w.r.t. unfoldable TBoxes, that have these formal properties. In this paper we investigate CSMs for $\mathcal{EL}$-concepts defined w.r.t. general TBoxes. We extend the definition of the properties of CSMs from (Lehmann and Turhan 2012) to the case of general TBoxes.

Definition 6. A CSM $\sim : \mathfrak{C}(\mathcal{EL}) \times \mathfrak{C}(\mathcal{EL}) \to [0, 1]$ is:

symmetric iff $C \sim_{\mathcal{T}} D = D \sim_{\mathcal{T}} C$;

equivalence invariant iff for all $C \equiv_{\mathcal{T}} D$ it holds that $C \sim_{\mathcal{T}} E = D \sim_{\mathcal{T}} E$;

equivalence closed iff $C \equiv_{\mathcal{T}} D \iff C \sim_{\mathcal{T}} D = 1$;

bounded iff the existence of $E \neq \top$ with $C \sqsubseteq_{\mathcal{T}} E$ and $D \sqsubseteq_{\mathcal{T}} E$ implies $C \sim_{\mathcal{T}} D > 0$;

dissimilar closed iff $C, D \neq \top$ and there is no $E \neq \top$ with $C \sqsubseteq_{\mathcal{T}} E$ and $D \sqsubseteq_{\mathcal{T}} E$ implies $C \sim_{\mathcal{T}} D = 0$;

subsumption preserving iff $C \sqsubseteq_{\mathcal{T}} D \sqsubseteq_{\mathcal{T}} E$ implies $C \sim_{\mathcal{T}} D \geq C \sim_{\mathcal{T}} E$;

reverse subsumption preserving iff $C \sqsubseteq_{\mathcal{T}} D \sqsubseteq_{\mathcal{T}} E$ implies $D \sim_{\mathcal{T}} E \geq C \sim_{\mathcal{T}} E$, and

These formally defined properties make CSMs more predictable for users. The measures in (Suntisrivaraporn 2013; Lehmann and Turhan 2012) fulfill most of these properties, in particular, they are symmetric and equivalence invariant, two properties that we will require later to compute relaxed instances. The parameterizable similarity measures from (Lehmann and Turhan 2012) additionally allow users to calibrate the measure to fit their expectations. In our setting these parametrizable CSMs enable users to specify which features of query concepts should be relaxed.

3 Relaxed Instances

As discussed before, our goal is to generalize query answering to allow for relaxing the solution set. This means, when a query concept Q is given, we are not only interested in the instances of Q, but also those individuals that are close enough to being instances; those individuals are called *relaxed instances*.

There are many ways to formalized the notion of relaxed instances, but in this paper we make use of CSMs for this relaxation. We say that relaxed instances of the query concept are those individuals that are instance of a (possibly different) concept that is sufficiently similar to the original query concept. The condition of 'sufficiently similar' is expressed via a threshold, or minimal similarity, t. One can then control how inclusive the relaxed instance solutions should be, by adjusting t, and guide the direction, in which solutions are relaxed by choosing an appropriate CSM.

Definition 7 (relaxed instance)**.** Let $\mathcal{L}$ be some DL, $\sim$ be a CSM, and $t \in [0, 1)$. The individual $a \in N_I$ is a *relaxed instance* of the query concept Q w.r.t. the $\mathcal{L}$-knowledge base $\mathcal{K}$, $\sim$ and the threshold t iff there exists a $\mathcal{L}$-concept description $X \in \mathfrak{C}(\mathcal{L})$ such that $Q \sim X > t$ and $\mathcal{K} \models X(a)$. $\text{Relax}_t^{\sim}(Q)$ denotes the set of all individuals occurring in $\mathcal{K}$ that are relaxed instances of Q w.r.t. $\mathcal{K}$, $\sim$ and t.

The task we are mostly interested in is not *relaxed instance checking*, i.e. testing whether a given individual a is a relaxed instance of the query concept Q, but to compute all relaxed instances of Q, i.e. the set $\text{Relax}_t^{\sim}(Q)$.

One naive idea to compute this set would be to first computing all concepts X that are similar to C with degree greater than t, and then obtaining all the instances of these concepts X; in symbols,

$$\text{Relax}_t^{\sim}(C) = \bigcup_{C \sim X > t} \{a \mid a \text{ is an instance of } X\}.$$

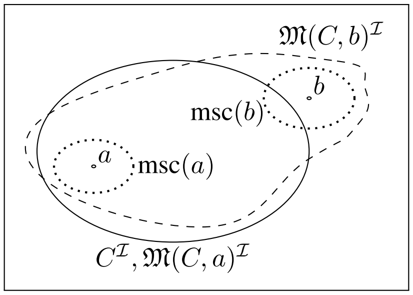
$\mathfrak{M}(C,b)^{\mathcal{I}}$
b
msc(b)
a
msc(a)
$C^{\mathcal{I}}, \mathfrak{M}(C,a)^{\mathcal{I}}$

Figure 2: Two individuals, their most specific concepts (dotted), and the mimics of a concept C w.r.t. the individuals (dashed).

However, this approach is not feasible, since there might be an infinite number of concepts X sufficiently similar to C. Furthermore, given an CSM, it is not apparent how to even obtain all concepts that are similar to C with degree $> t$, since in general CSMs only allow to evaluate the similarity of two given concepts.

To avoid these issues, we consider a different reasoning problem, that considers the computation of a concept that has a given individual a as an instance and resembles C best w.r.t. the given CSM. We call this the *mimic* of C w.r.t. a.

Definition 8 (mimic)**.** Let $\mathcal{L}$ be a DL, $\mathcal{K}$ be an $\mathcal{L}$-knowledge base, $a \in N_I$ be an individual name, C be an $\mathcal{L}$-concept description, and $\sim$ be a CSM. An $\mathcal{L}$-concept D is called a *mimic* of C w.r.t. a, denoted $D \in \mathfrak{M}(C, a)$, iff the following two conditions hold:

- a is an instance of D, i.e., $a^{\mathcal{I}} \in D^{\mathcal{I}}$ for all models $\mathcal{I}$ of $\mathcal{K}$, and
- for all $\mathcal{L}$-concept descriptions E holds, if a is an instance of E, then $C \sim D \geq C \sim E$.

Intuitively, a mimic of C w.r.t. a is the most similar concept to C that has a as an instance. In general, the mimic of C w.r.t. an individual a may not be unique, even modulo concept equivalence, however, for our purposes it suffices to compute only one of them.

Figure 2 depicts the idea of mimics. In the figure, a and b are two different individuals. Since a is an instance of C, C is also a mimic of C w.r.t. a because $C \sim C = 1$. The dashed line depicts a mimic of C w.r.t. b. Notice that this mimic must contain b and thus subsume msc(b), but need not be a subsumer of C, it just tries to resemble C in a way that results in maximal similarity.

There is an easy reduction from the computation of relaxed instances to the computation of mimics. The idea is that, for each individual a appearing in the knowledge base $\mathcal{K}$, the mimic of Q w.r.t. a must have a similarity greater than t to the query concept Q for a to be a relaxed instance of Q; if not, a cannot be a relaxed instance, as no concept can have a greater similarity to Q while containing a. This is formalized in the following proposition.

Proposition 9. *Let $\mathcal{K}$ be a knowledge base, a be an individual occurring in $\mathcal{K}$, Q be a concept description, $\sim$ be a*

CSM and $t \in [0,1]$. Then $a \in \mathrm{Relax}_t^{\sim}(Q)$ iff there is a mimic $D \in \mathfrak{M}(Q,a)$ of Q w.r.t. individual a such that $Q \sim D > t$.

This connection allows to study the problem of computing mimics in order to find relaxed instances. In the next section, we will show how this can be solved for CSMs on unfoldable terminologies.

4 Computing Relaxed Instances w.r.t. Unfoldable $\mathcal{EL}$-Terminologies

Structural concept similarity measures with nice properties are often only defined for unfoldable TBoxes, as in (Lehmann and Turhan 2012) and (Suntisrivaraporn 2013). The reason is that in those terminologies, any concept can be expanded by exhaustively replacing all defined concept names by their definitions, until only atomic concept names remain; the concept is then called *fully expanded*. After expanding all concepts, the TBox can be completely disregarded for computing similarities or other inferences like subsumption. In this section, we show how to compute mimics, and therefore relaxed instances w.r.t. unfoldable $\mathcal{EL}$-terminologies.

In general there may exist infinitely many concept descriptions which have the individual a as an instance, and thus enumerating them and finding the maximal similarity to the query concept Q to compute the mimic is not a feasible option. However, under some circumstances we can limit the number of concepts that need to be tested.

Recall that any mimic $D \in \mathfrak{M}(Q,a)$ must have a as an instance, and thus, by definition of the msc, $\mathrm{msc}(a) \sqsubseteq_{\mathcal{T}} D$ holds. For equivalence invariant similarity measures one can use the $\mathrm{msc}(a)$ as a lower bound for the mimic, and only consider concept descriptions that can be obtained from syntactic manipulations of $\mathrm{msc}(a)$ that result in a generalized concept, i.e., by removing some concept names or existential restrictions.

Definition 10 (generalized concept)**.** Let C be a concept description of the form

$$C = \textstyle\bigsqcap_{i\in I} A_i \ \sqcap\ \bigsqcap_{j\in J} \exists r_j.E_j,$$

with $A_i \in N_C$ for all $i \in I$, and $r_j \in N_R$, E_j is a concept description for all $j \in J$. Then a concept description D is a *generalized concept* of C iff it has the form

$$D = \textstyle\bigsqcap_{i\in I'} A_i \ \sqcap\ \bigsqcap_{j\in J'} \exists r_j.E'_j$$

with $I' \subseteq I$, $J' \subseteq J$ and E'_j is a generalized concept of E_j for $j \in J'$.

Clearly, if D is a generalized concept of C, then we always have $C \sqsubseteq D$, since generalizing only removes parts of C, but never adds new features. Indeed, in our setting the other direction holds as well for unfoldable TBoxes $\mathcal{T}$: If $\mathrm{msc}(a)$ is fully expanded w.r.t. $\mathcal{T}$, then any concept description D with $\mathrm{msc}(a) \sqsubseteq D$ is equivalent to a generalized concept of $\mathrm{msc}(a)$. If we now restrict to equivalence invariant CSMs, we get the following result for the mimic.

Lemma 11. *Let $\mathcal{K} = (\mathcal{T},\mathcal{A})$ be an unfoldable $\mathcal{EL}$-KB, a be an individual from $\mathcal{A}$, C be an $\mathcal{EL}$-concept, and $\sim$ be an equivalence invariant similarity measure. If $E = \mathrm{msc}(a)$ is the fully expanded most specific concept of a, then there is a mimic $D \in \mathfrak{M}(C,a)$ of C w.r.t. a and $\mathcal{K}$ that is a generalized concept of E.*

Proof sketch. We show that any concept having a as an instance is equivalent to a generalized concept of $\mathrm{msc}(a)$. Let F be a concept description with $\mathcal{K} \models F(a)$. Since $\sim$ is equivalence invariant, we can assume F to be fully expanded as well. $E \sqsubseteq_{\mathcal{K}} F$ holds by definition of the msc. As both E and F are fully expanded and thus do not contain defined concept names from the TBox, any subconcept of F must also be a subconcept E, but F may contain redundancies that can simply be removed to get an equivalent concept. Thus F is equivalent to a generalized concept of E. □

However, in general the msc may result in infinitely nested existential restrictions, and thus could only be described by concept descriptions of infinite size. In this case there are infinitely many generalized concepts of finite size that would need to be checked to find a mimic. Lemma 11 therefore does not always provide a solution to the problem. However, the query concept Q has always a finite role-depth, even if fully expanded. Since most structural similarity measures used in practice, like those presented in (Suntisrivaraporn 2013; Lehmann and Turhan 2012), compute the similarity recursively between sub-concepts at the same role-depth, it is possible to also limit the role-depth of the most specific concept and still get the same result:

Lemma 12. *Let $\mathcal{K} = (\mathcal{T},\mathcal{A})$ be an unfoldable $\mathcal{EL}$-KB, a be an individual from $\mathcal{A}$, Q be a fully expanded $\mathcal{EL}$-concept, and $\sim$ be a equivalence invariant similarity measure, which is additionally structural and depth-bounded, i.e.:*

$$C \sim D \geq C \sim E \Rightarrow \exists r.C \sim \exists r.D \geq \exists r.C \sim \exists r.E, \quad (1)$$

$$C \sim \textstyle\bigsqcap_{i\in I} A_i \ \geq\ C \sqcap \exists r.D \sim \bigsqcap_{i\in I} A_i, \quad (2)$$

for all fully expanded concepts C, D, and E, and atomic concept names A_i. If $k = \mathrm{rd}(Q)$ and $E = k$-$\mathrm{msc}(a)$ is the fully expanded most specific concept of a bounded to the role-depth k, then there is a mimic $D = \mathfrak{M}(Q,a)$ of Q w.r.t. a that is a generalized concept of E.

Proof sketch. Using the properties (1) and (2) above, we can show the following: Trimming the role-depth of two concepts C and D to the minimal role-depth $\min(\mathrm{rd}(C), \mathrm{rd}(D))$ never reduces the similarity value, i.e., the similarity value between the trimmed concepts must be greater than or equal to the similarity value of C and D.

Using this, we can prove the lemma as follows: Let M be any mimic of Q w.r.t. a. Since the mimic is always equivalent to a generalized concept M' of the (possibly infinite) $\mathrm{msc}(a)$, and since $\sim$ is equivalence invariant, M' is also a mimic. Using the previous property, we know that M' trimmed to role-depth k must still have (at least) the same similarity to Q, and hence also be a mimic. But M' trimmed to role-depth k is also a generalized concept of the k-$\mathrm{msc}(a)$. □

This shows that we can always find the mimic of Q w.r.t. a from a finite set of concept descriptions: the generalized

```
**Procedure:** relaxed-instance? $(a, Q, \mathcal{K}, \sim, t)$
**Input:** $a$: individual in $\mathcal{K}$; $Q$: $\mathcal{EL}$-concept; $\mathcal{K}$: unfoldable
    $\mathcal{EL}$-KB; $\sim$: CSM; $t \in (0, 1]$: threshold
**Output:** whether $a \in \text{Relax}_t^{\sim}(Q)$ w.r.t. $\mathcal{K}$
  $k := \text{rd}(Q)$
  $E := k\text{-msc}(a)$ w.r.t. $\mathcal{K}$
  guess a generalized concept $F$ of $E$
  **return** $F \sim Q > t$
```

Figure 3: Non-deterministic algorithm for relaxed instances checking w.r.t. unfoldable $\mathcal{EL}$ terminologies.

concepts of the fully expanded k-msc(a). However, instead of computing the mimic $D \in \mathfrak{M}(Q, a)$ and testing whether the similarity between the Q and D is greater than t, it is enough to find *any* concept D' with a as an instance and $Q \sim D' > t$ to show that a is a relaxed instance of Q; this gives rise to the non-deterministic algorithm in Figure 3, that checks whether a is a relaxed instance of Q.

Corollary 13. *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be an unfoldable $\mathcal{EL}$-KB, Q be a fully expanded $\mathcal{EL}$-concept, a be an individual in $\mathcal{K}$, $\sim$ be an equivalence invariant similarity measure fulfilling properties 1 and 2 from Lemma 12 and $t \in [0, 1]$. Then* relaxed-instance?$(a, Q, \mathcal{K}, \sim, t)$ *computes whether $a \in \text{Relax}_t^{\sim}(Q)$ w.r.t. $\mathcal{K}$.*

Guessing a generalized concept F of a concept description E can be done in time linear on the size $\|E\|$ of E by recursively guessing for each concept name and each existential restriction in E whether they should occur in F or not. However, the size of $E = k$-msc(a) can be exponential in k and polynomial in $\|\mathcal{K}\|$ (Peñaloza and Turhan 2011). Since $k = \text{rd}(C)$ is bounded linearly by $\|C\|$, the algorithm runs in NEXP-time (provided that $\sim$ can be computed in NEXP-time). However, the algorithm runs in NP-time in $\|\mathcal{K}\|$ (provided that $\sim$ can be computed in NP), and since C is an input concept, its role-depth can be assumed to be rather low. Hence, we conjecture that the exponential blow-up of the msc usually plays only a minor role in practical applications.

5 Relaxed Instances w.r.t. General $\mathcal{EL}$-TBoxes

The approach presented in the previous section does not work for general $\mathcal{EL}$-TBoxes since those structural similarity measures first fully expand the concepts w.r.t. the background knowledge. As soon as the background knowledge contains cyclic concept inclusions, this expansion is no longer possible. In fact, to the best of our knowledge no equivalence invariant, structural concept similarity measure that can handle general $\mathcal{EL}$-TBoxes has been introduced in the literature so far. In this section, we therefore define such a concept similarity measure $\sim_c$; this measure can then be used to find all relaxed instances for such TBoxes.

The approach to compute the similarity $C \sim_c D$ w.r.t. $\mathcal{T}$ is to translate the concepts C and D into their canonical models $\mathcal{I}_{C,\mathcal{T}}$ and $\mathcal{I}_{D,\mathcal{T}}$ and then structurally compute the similarities for the elements d_C and d_D in these interpretations. As observed before, $\mathcal{I}_{C,\mathcal{T}}$ and $\mathcal{I}_{D,\mathcal{T}}$ are always finite. The TBox $\mathcal{T}$ itself can be disregarded when computing the interpretation similarity from the canonical models, since all its axioms are already used to construct the models.

We need to define similarity measures on interpretations and their properties, which are derived from the properties of concept similarity measures. All proofs for this section can be found in the technical report (Ecke and Turhan 2013).

Interpretation Similarity Measures

An *interpretation similarity measure* (ISM) is defined as a similarity measure on pointed interpretations, i.e., a function $\sim_{\mathfrak{P}} : \mathfrak{P} \times \mathfrak{P} \to [0, 1]$ such that $p \sim_{\mathfrak{P}} p = 1$ for all $p \in \mathfrak{P}$. It maps any pair of pointed interpretations to a similarity value between 0 and 1.

There are various desirable properties that ISMs can have. We concentrate here on those that directly transfer from analogous properties of CSMs introduced by (Lehmann and Turhan 2012). Given suitable simulation relations $\lesssim$ and $\simeq$ (like e.g., those defined in Definition 1 for $\mathcal{EL}$), we call an interpretation similarity measure:

- *symmetric* iff $p \sim_{\mathfrak{P}} q = q \sim_{\mathfrak{P}} p$ for all $p, q \in \mathfrak{P}$;
- *bounded* iff $\mathfrak{C}(p) \cap \mathfrak{C}(q) \supsetneq \{\top\}$ implies $p \sim_{\mathfrak{P}} q > 0$ for all $p, q \in \mathfrak{P}$;
- *dissimilar closed* iff $\mathfrak{C}(p) \cap \mathfrak{C}(q) = \{\top\}$ implies $p \sim_{\mathfrak{P}} q = 0$ for all $p, q \in \mathfrak{P}$ with $\mathfrak{C}(p) \supsetneq \{\top\}$ and $\mathfrak{C}(q) \supsetneq \{\top\}$;
- *equisimulation invariant* iff for all $p, q, u \in \mathfrak{P}$, $p \simeq q$ implies $p \sim_{\mathfrak{P}} u = q \sim_{\mathfrak{P}} u$;
- *equisimulation closed* iff $p \simeq q \iff p \sim_{\mathfrak{P}} q = 1$ for all $p, q \in \mathfrak{P}$;
- *simulation preserving* iff for all $p, q, r \in \mathfrak{P}$, $r \lesssim q \lesssim p$ implies $p \sim_{\mathfrak{P}} q \geq p \sim_{\mathfrak{P}} r$;
- *reverse simulation preserving* iff $r \lesssim q \lesssim p$ implies $q \sim_{\mathfrak{P}} r \geq p \sim_{\mathfrak{P}} r$ for all $p, q, r \in \mathfrak{P}$.

We now define a parameterizable ISM $\sim_i$, using the simulation relations defined in Definition 1, which correspond to subsumption and equivalence in $\mathcal{EL}$. This is important when lifting those properties to the CSMs $\sim_c$. Given a pointed interpretation $p = (\mathcal{I}, d)$, we denote with

- $\text{CN}(p) = \{A \in N_C \mid d \in A^{\mathcal{I}}\}$ the set of concept names that have d as instance in $\mathcal{I}$, and
- $\text{SC}(p) = \{(r, (\mathcal{I}, e)) \in N_R \times \mathfrak{P} \mid (d, e) \in r^{\mathcal{I}}\}$ the set of direct successors of d in $\mathcal{I}$.

For two pointed interpretations to be perfectly similar, they need to have the same set of concept names and edges labeled with the same roles going to perfectly similar successor elements. Otherwise, the most similar concept names and the most similar direct successors are compared and a similarity value is computed from this pair. For both cases, we need the notion of pairings:

A *pairing* $P \subseteq X \times Y$ is a total binary relation, where totality means that all elements of X and all elements of Y appear in some tuple of P as the first component or second component, respectively. For two pointed interpretations p and q, we are interested in two types of pairings:

- $P_C(p,q) \subseteq \mathcal{P}(\text{CN}(p) \times \text{CN}(q))$ is the set of all *concept name pairing* on the concept names that p and q are instances of; and
- $P_S(p,q) \subseteq \mathcal{P}(\text{SC}(p) \times \text{SC}(q))$ is the set of all *successor pairings* on the direct successors of p and q.

Note that when one of CN(p) or CN(q) is empty, the pairings can not be total. In this case, we use $\{\top\}$ instead of the empty concept name set. Similarly, when one of SC(p) or SC(q) is empty, we instead use the set $\{(r_\top, p)\}$ or $\{(r_\top, q)\}$, respectively, where $r_\top \in N_R$ is a new role name. This way, the pairings are always well-defined.

The ISM $\sim_i$ will extend a primitive measure. A *primitive measure* $\sim_{\text{prim}} : N_C \cup \{\top\} \times N_C \cup \{\top\} \cup N_R \times N_R \to [0,1]$ assigns similarity values to each pair of basic concepts (i.e., concept names or $\top$) and each pair of role names. Any primitive measure has to satisfy the following properties: $x \sim_{\text{prim}} x = 1$ for any role name or basic concept x, $\top \sim_{\text{prim}} A = A \sim_{\text{prim}} \top = 0$ for all $A \in N_C$, and similarly $r_\top \sim_{\text{prim}} s = s \sim_{\text{prim}} r_\top = 0$ for all $s \in N_R \setminus \{r_\top\}$. Additionally, for the similarity measure $\sim_i$ to be symmetric, $\sim_{\text{prim}}$ needs to be symmetric as well.

We give a default primitive measure, that simply assigns similarity 0 to pairs of different basic concepts or role names:

$$x \sim_{\text{default}} y = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

However, other primitive measures are imaginable and useful. For example, one might want to express that two colours Red and Orange are similar to *some* degree even if they have different concept names. Additionally, one can assign weights to different concept and role names through a weighting function $g : N_C \cup N_R \to \mathbb{R}_{>0}$ that prioritizes different features of the similarity measure. This function g can be extended to pairs of concept names as $g(A,B) = \max(g(A), g(B))$ and pairs of role names as $g(r,s) = \max(g(r), g(s))$.

Any primitive measure $\sim_{\text{prim}}$ and weighting function g can then be extended to a similarity measure on pointed interpretations by recursively traversing the interpretation graphs, computing the primitive measure for each pair of concepts in the best concept name pairing and combining it with the discounted similarity between all pairs of successors in the best successor pairing at each element.

Definition 14. Given a primitive measure $\sim_{\text{prim}}$, a weighting function g, and a discounting factor $w \in (0,1)$, the ISM $\sim_i$: $\mathfrak{P} \times \mathfrak{P} \to [0,1]$ is defined as follows:

$$p \sim_i q = \max_{\substack{P_C \in P_C(p,q) \\ P_S \in P_S(p,q)}} \left(\frac{\text{sim}(P_C) + \text{sim}(P_S)}{\sum\limits_{(A,B)\in P_C} g(A,B) + \sum\limits_{((r,p'),(s,q'))\in P_S} g(r,s)} \right)$$

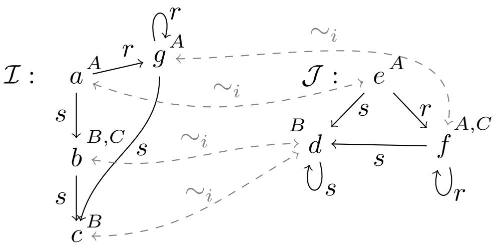

Figure 4: Computation of $(\mathcal{I}, a) \sim_i (\mathcal{J}, e)$ by pairing their successors and recursively computing $\sim_i$ between those.

where

$$\text{sim}(P_C) = \sum_{(A,B)\in P_C} g(A,B)(A \sim_{\text{prim}} B), \text{ and}$$

$$\text{sim}(P_S) = \sum_{\substack{((r,p'),(s,q')) \\ \in P_S}} g(r,s)(r \sim_{\text{prim}} s)((1-w) + w(p' \sim_i q')).$$

The constant w allows for discounting of successors, and should have a value $0 < w < 1$.

Example 15. Consider the pointed interpretations $(\mathcal{I}, a)$ and $(\mathcal{J}, e)$ depicted in Figure 4, the default primitive measure, a weighting function g that assigns 1 to all concept and role names, and a discounting factor $w = 0.8$. To compute the similarity $(\mathcal{I}, a) \sim_i (\mathcal{J}, e)$, we proceed as follows.

1. To compute the similarity between c and d, we find the only concept name pairing is $\{(B,B)\}$ and the only successor pairing is $\{((r_\top, c), (s,d))\}$. Thus we have that $(\mathcal{I}, a) \sim_i (\mathcal{J}, b)$ is $\frac{B \sim_{\text{default}} B + (r_\top \sim_{\text{default}} s) \cdot \ldots}{2} = 0.5$.
2. The similarity between b and d, for the best concept name pairing $\{(B,B),(C,B)\}$ and the best successor pairing $\{((s,c),(s,d))\}$, is $\frac{1+0+(0.2+0.8 \cdot 0.5)}{3} = 0.533$.
3. The similarity between g and f, for the best concept name pairing $\{(A,A),(A,C)\}$ and the best successor pairing $\{((s,c),(s,d)),((r,g),(r,f))\}$, is the solution to the equation $x = \frac{1+0+(0.2+0.8 \cdot 0.5)+(0.2+0.8 \cdot x)}{4}$, which is $x = 0.563$.
4. Finally, the similarity between a and e, for the best concept name pairing $\{(A,A)\}$ and the best successor pairing $\{((s,b),(s,d)),((r,g),(r,f))\}$, is $\frac{1+(0.2+0.8 \cdot 0.533)+(0.2+0.8 \cdot 0.563)}{3} = 0.759$.

If role s is very important for the similarity, one can adapt the similarity measure by increase the weight $g(s)$. Similarly, if concept name A is less important than the other concept names, since it only describes a broad category, one can decrease its weight. One can also change the primitive measure to evaluate $r \sim_{\text{prim}} s = 0.5$ if roles r and s actually denote similar relations, even though they use different role names. In general, this might imply that different best pairings will be found. Finally, if one wants to put more emphasize on the similarity of the role names itself and less on the similarity of the actual elements of the successors, one can do so by reducing the discounting factor. These three degrees of freedom allow for the adaptation of $\sim_i$ to many different use cases.

Note that by defining the ISM $\sim_i$ this way, it is not equisimulation closed. The reason is that the successor pairing always connects successors symmetrically, which gives rise to problems for the case where one successor of an element simulates a second successor. To regain equivalence invariance, one can first normalize the interpretations $\mathcal{I}$ and $\mathcal{J}$ before applying the similarity measure.

Definition 16 (normal form for interpretations)**.** An interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ is in *normal form* if there are no elements $a, b, c \in \Delta^{\mathcal{I}}$ with $\{(a, b), (a, c)\} \in r^{\mathcal{I}}$ and $(\mathcal{I}, b) \lesssim (\mathcal{I}, c)$, i.e., no node has two successor nodes for the same role name that are in a simulation relation.

Any interpretation $\mathcal{I}$ can be transformed into normal form as follows:

1. remove all edges $(a, b) \in r^{\mathcal{I}}$ in the interpretation graph, for which there exists an edge $(a, c) \in r^{\mathcal{I}}$ such that $(\mathcal{I}, b) \lesssim (\mathcal{I}, c)$ but not $(\mathcal{I}, b) \simeq (\mathcal{I}, c)$
2. for all edges $(a, b_0) \in r^{\mathcal{I}}$, check if there are other edges $(a, b_i) \in r^{\mathcal{I}}$, $i > 0$, with $(\mathcal{I}, b_0) \simeq (\mathcal{I}, b_i)$ and choose one representative b_j; then remove all other edges (a, b_i), $i \neq j$, from $r^{\mathcal{I}}$.

Note that this normalization is well-defined, since the pointed interpretations are always finite, and simulations can be computed in polynomial time in the size of the interpretation. In the following, whenever we write $(\mathcal{I}, a) \sim_i (\mathcal{J}, b)$, we implicitly assume that $\mathcal{I}$ and $\mathcal{J}$ have been normalized first. Using this, we can finally show that $\sim_i$ is a well-defined ISM with formal properties:

Theorem 17. *The similarity measure $\sim_i$ is well-defined, i.e., $p \sim_i q$ has a unique solution for all pointed interpretations $p, q \in \mathfrak{P}$. Furthermore $\sim_i$ is symmetric, bounded, dissimilar closed, equisimulation invariant, and equisimulation closed for the simulation relations defined in Definition 1 and normalized pointed interpretations.*

Although $\sim_i$ as given in Definition 14 is well-defined, it cannot be used directly to compute the similarity value, since cycles in the interpretation would lead to infinite recursion. Instead, one can view the defining equation as an iterative algorithm: When starting with a similarity value of 0 between all elements of two interpretations $\mathcal{I}$ and $\mathcal{J}$, and iteratively applying the equation to update those similarity values, they will converge to the solution in the limit. This follows from the Banach fixed-point theorem (Banach 1922), as $\sim_i$ can be seen as a contraction mapping on the similarity values between all elements of $\mathcal{I}$ and $\mathcal{J}$.

Using this canonical model, we define a concept similarity measure $\sim_c$ on $\mathcal{EL}$-concept descriptions w.r.t. a general $\mathcal{EL}$-TBox $\mathcal{T}$ as follows:

$$C \sim_c D = (\mathcal{I}'_{C,\mathcal{T}}, d_C) \sim_i (\mathcal{I}'_{D,\mathcal{T}}, d_D),$$

where $\mathcal{I}'_{C,\mathcal{T}}$ and $\mathcal{I}'_{D,\mathcal{T}}$ are the normalized canonical models of C and D w.r.t. $\mathcal{T}$.

The concept similarity measure $\sim_c$ inherits the useful properties of $\sim_i$, since the properties for interpretation similarity measures were defined to correspond exactly to the concept similarity properties given in the preliminaries.

Theorem 18 (Properties of $\sim_c$)**.** *The concept similarity measure $\sim_c$ is symmetric, bounded, dissimilar closed, equivalence invariant, and equivalence closed.*

The proof for this theorem is given in (Ecke and Turhan 2013).

Computing Relaxed Instances w.r.t. $\sim_c$

First we define the notion of *fully expanded concepts* also for the general $\mathcal{EL}$ case:

Definition 19 (fully expanded concept)**.** Let $\mathcal{T}$ be a general $\mathcal{EL}$-TBox. A concept description C is *fully expanded* w.r.t. $\mathcal{T}$ iff for all GCIs $D \sqsubseteq E \in \mathcal{T}$ with $C \sqsubseteq_{\mathcal{T}} \exists r_1 \ldots \exists r_n.D$ we have that $\exists r_1 \ldots \exists r_n.E$ is a generalized concept of C.

For the computation of relaxed instances for $\sim_c$, recall that $a \in \mathit{Relax}_t^{\sim}(Q)$ can be computed for unfoldable terminologies by checking all generalized concepts of the k-msc(a) for $k = \text{rd}(Q)$. Of course, as soon as we have a general TBox, expanding Q may result in an infinite role-depth by expanding cyclic definitions, so this approach does not work directly here. However, given a KB $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, one can express the fully expanded msc(a) in $\mathcal{EL}$ by unraveling the canonical model $\mathcal{I}_{\mathcal{K}}$ starting from d_a, and thus for any concept C we have

$$C \sim_c \text{msc}(a) = (\mathcal{I}_{C,\mathcal{T}}{}', d_C) \sim_i (\mathcal{I}_{\mathcal{K}}{}', d_a),$$

where $\mathcal{I}_{C,\mathcal{T}}{}'$ and $\mathcal{I}_{\mathcal{K}}{}'$ are the normalized canonical models of C and $\mathcal{A}$ w.r.t. the TBox $\mathcal{T}$. The canonical model $\mathcal{I}_{\mathcal{K}}$, in contrast to the fully expanded msc, is always finite.

However, we do not need to compute the similarity between the query concept Q and the msc(a) directly, but find the maximal similarity between Q and generalized concepts of msc(a). Generalizing a concept C is possible by removing concept names or existential restriction, which corresponds on the interpretation side to only taking subsets of the concept names $S_{\text{CN}} \subseteq \text{CN}(q)$ and successors $S_{\text{SC}} \subseteq \text{SC}(q)$ of the pointed interpretations $q = (\mathcal{I}_{C,\mathcal{T}}, d_C)$ and all its successors. This gives the algorithm in Figure 5 to iteratively compute the maximal similarity between a pointed interpretation p and all generalizations of the pointed interpretation q.

Using this, the algorithm to actually compute all relaxed instances of a query concept Q w.r.t. $\sim_c$ is conceptually quite easy, as it only needs to compute the maximal similarities between Q and all individuals a and check whether they are larger than t. The algorithm is depicted in Figure 6.

Example 20. Consider the the KB $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ consisting of the following TBox $\mathcal{T}$ and knowledge base $\mathcal{A}$:

$$\begin{aligned}\mathcal{T} &= \{\, A \sqsubseteq \exists r.A \sqcap \exists s.B,\;\; C \sqsubseteq \exists s.B \,\},\\ \mathcal{A} &= \{\, A(e),\;\; A \sqcap C(f),\;\; B(d),\\ &\qquad r(e, f),\;\; r(f, f),\;\; s(f, d),\\ &\qquad s(e, d),\;\; s(d, d) \,\}.\end{aligned}$$

The similarity measure used is the same as in Example 15, with the default weighting function, the discounting factor $w = 0.8$ and the default primitive measure. Given the query concept $Q = A \sqcap \exists s.(B \sqcap C)$, the normalized canonical

```
Procedure: maxsim $(\mathcal{I}, \mathcal{J}, \sim_{\text{prim}}, g, w)$
Input: $\mathcal{I}, \mathcal{J}$: finite interpretations; $\sim_{\text{prim}}$: primitive measure; $g$: weighting function; $w \in (0,1)$: discount factor
Output: maximal similarities between $p = (\mathcal{I}, a)$ and all generalizations of $q = (\mathcal{J}, b)$
msim$_0(d,e) \leftarrow 0$ for all $d \in \Delta^{\mathcal{I}}$ and $e \in \Delta^{\mathcal{J}}$
for $i \leftarrow 1$ to $n$ do
    for all $d \in \Delta^{\mathcal{I}}$ and $e \in \Delta^{\mathcal{J}}$ do
        $\text{msim}_i(d,e) \leftarrow \max_{\substack{S_{\text{CN}} \subseteq \text{CN}(e) \\ S_{\text{SC}} \subseteq \text{SC}(e)}} \Big( \max_{\substack{P_C \subseteq \text{CN}(d) \times S_{\text{CN}} \\ P_S \subseteq \text{SC}(d) \times S_{\text{SC}}}} \text{similarity}(P_C, P_S, \sim_{\text{prim}}, g, w, i) \Big)$
    end for
end for

Procedure: similarity $(P_C, P_S, \sim_{\text{prim}}, g, w, i)$
sim$(P_C) \leftarrow \sum_{(A,B) \in P_C} g(A,B)(A \sim_{\text{prim}} B)$
sim$(P_S) \leftarrow \sum_{((r,p'),(s,q')) \in P_S} g(r,s)(r \sim_{\text{prim}} s)\big((1-w) + w \cdot \text{msim}_{i-1}(p',q)\big)$
return $\dfrac{\text{sim}(P_C) + \text{sim}(P_S)}{\sum_{(A,B) \in P_C} g(A,B) + \sum_{((r,p'),(s,q')) \in P_S} g(r,s)}$
```

Figure 5: Algorithm to compute the maximal similarities between all elements of two finite interpretations $\mathcal{I}$ and $\mathcal{J}$.

models $\mathcal{I}'_{Q,\mathcal{T}}$ and $\mathcal{I}'_{\mathcal{K}}$ are exactly the interpretations $\mathcal{I}$ and $\mathcal{J}$ found in Example 15, with $d_Q = a$. The algorithm maxsim will compute the following similarity values msim$_i$ between d_Q and d, e, f in each iteration i:

i	msim$_i(d_Q, d)$	msim$_i(d_Q, e)$	msim$_i(d_Q, f)$
0	0	0	0
1	0.067	0.467	0.35
2	0.173	0.667	0.5
3	0.209	0.748	0.561
4	0.209	0.757	0.567
5	0.209	0.758	0.569

Since the both canonical models are cyclic, the similarity values will change slightly in further iterations, but already after the 5th, the maximal error is less than 0.05%. Using these values, we see that for the threshold $t = 0.5$, both e and f are relaxed instance of Q, while d is not.

```
Procedure: relaxed-instances $(Q, \mathcal{K}, t, \sim_{\text{prim}}, g, w)$
Input: $Q$: $\mathcal{EL}$-concept; $\mathcal{K} = (\mathcal{T}, \mathcal{A})$: $\mathcal{EL}$-KB; $t \in [0,1]$: threshold; $\sim_{\text{prim}}$: primitive measure; $g$: weighting function; $w \in (0,1)$: discounting factor
Output: individuals $a \in Relax_t^{\sim_c}(Q)$
compute canonical models $\mathcal{I}_{Q,\mathcal{T}}$ and $\mathcal{I}_{\mathcal{K}}$
maxsim$(d,e) \leftarrow$ maxsim$(\mathcal{I}_{Q,\mathcal{T}}, \mathcal{I}_{\mathcal{K}}, \sim_{\text{prim}}, g, w)$
return $\{a \in N_I \cap \text{Sig}(\mathcal{A}) \mid \text{maxsim}(d_Q, d_a) > t\}$
```

Figure 6: Algorithm to compute all relaxed instances of a query concept Q w.r.t. a knowledge base $\mathcal{K}$ and threshold t.

The maxsim$_i$ values computed in the algorithm monotonically converge from below to the maximal similarities between generalized concepts of the most specific concept of an individual and the query concept. Thus, for any individual a, which is a relaxed instance of Q with a threshold strictly larger than t, there exists $i \in \mathbb{N}$ such that for all $j > i$ we have maxsim$_j(Q, a) > t$. Thus, the algorithm is sound and complete in the following sense:

Theorem 21. *Let $\sim_c$ be the CSM derived from $\sim_i$ with the primitive measure $\sim_{\text{prim}}$, the weighting function g and the discounting factor w. Then the algorithm* relaxed-instances *is sound and complete:*

1. Soundness*: If $a \in$ relaxed-instances$(Q, \mathcal{K}, t, \sim_{\text{prim}}, g, w)$ for a number n of iterations, then $a \in Relax_t^{\sim_c}(Q)$.*
2. Completeness*: If $a \in Relax_t^{\sim_c}(Q)$, then there exists an $i \in \mathbb{N}$ such that for all $n \geq i$ iterations it holds that $a \in$ relaxed-instances$(Q, \mathcal{K}, t, \sim_{\text{prim}}, g, w)$.*

Furthermore, the algorithm converges quite fast: For any iteration, the difference between the actual similarity and the computed value reduces by a factor of w. This is again a direct consequence of the Banach fixed-point theorem, as w is an upper bound for the Lipschitz constant of the contraction mapping. This means that, to reduce the error tolerance of the solutions by a constant factor, e.g. one tenth, only a constant number of iterations need to be done additionally. However, one cannot compute how many iterations are needed beforehand and cannot be sure if, at any given point, the algorithm already found all relaxed instances, or if some relaxed instances with a maximal similarity very close to the threshold t are still missing.

If applying the algorithm relaxed-instances w.r.t. unfoldable TBoxes $\mathcal{T}$, then maxsim will however return the exact answer after exactly k iterations, where $k = \text{rd}(Q) + 1$ is the role-depth of the query concept Q expanded w.r.t. $\mathcal{T}$. In this case, the algorithm can be made deterministic and, since each iteration of maxsim only takes polynomial time in the size of $\mathcal{K}$ and Q, runs in PTIME.

6 Conclusions

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. The inference has two main degrees of freedom: in the choice of the CSM, and in the degree of relaxation of the concept. Intuitively, different similarity measures yield different weights on specific criteria. For example, one could require that small changes inside existential restrictions produce a high level of dissimilarity.

Further 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 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 for general TBoxes. Based on these we gave an computation algorithm for relaxed instances w.r.t. general TBoxes.

There are many options for future work. On the theoretical side it would be interesting to explore how this approach can be extended to expressive DLs. We conjecture that our approach extends to Horn-DLs, since they induce canonical models as well. On the practical side there is plenty of room for optimizations. For instance, the use of a concept that states necessary conditions in combination with the query concept can considerably reduce the number of individuals to be checked in practice. Furthermore, while the complexity of each iteration in the general case is polynomial, the need to check every subset and every pairing is certainly inefficient. Methods to reduce the subsets and pairings that need to be considered are expedient to make this work in practice.

Acknowledgements

A. Ecke is supported by the German Research Foundation (DFG) in the Graduiertenkolleg 1763 (QuantLA). R. Peñaloza is partially supported by DFG within the Cluster of Excellence 'Center for Advancing Electronics Dresden' (cfAED). A.-Y. Turhan is partially supported by DFG in the Collaborative Research Center 912 'Highly Adaptive Energy-Efficient Computing' (HAEC).

References

Alvarez, M. A., and Yan, C. 2011. A graph-based semantic similarity measure for the gene ontology. *J. Bioinformatics and Computational Biology* 9(6):681–695.

Banach, S. 1922. Sur les oprations dans les ensembles abstraits et leur application aux quations intgrales. *Fundamenta Mathematicae* 3(1):133–181.

Borgida, A.; Walsh, T.; and Hirsh, H. 2005. Towards measuring similarity in description logics. volume 147 of *CEUR Workshop Proceedings*.

Borgwardt, S., and Peñaloza, R. 2012. Undecidability of fuzzy description logics. In Brewka, G.; Eiter, T.; and McIlraith, S. A., eds., *KR-12*, 232–242. AAAI Press.

d'Amato, C.; Fanizzi, N.; and Esposito, F. 2005. A semantic similarity measure for expressive description logics. In *Proc. of Convegno Italiano di Logica Computazionale, CILC05*.

d'Amato, C.; Staab, S.; and Fanizzi, N. 2008. On the influence of description logics ontologies on conceptual similarity. In Gangemi, A., and Euzenat, J., eds., *Proceedings of Knowledge Engineering: Practice and Patterns, 16th Int. Conf. (EKAW 2008)*, volume 5268, 48–63.

Ecke, A., and Turhan, A.-Y. 2013. 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. See `http://lat.inf.tu-dresden.de/research/reports.html`.

Ecke, A.; Peñaloza, R.; and Turhan, A.-Y. 2013. Towards instance query answering for concepts relaxed by similarity measures. In Godo, L.; Prade, H.; and Qi, G., eds., *Workshop on Weighted Logics for AI (in conjunction with IJCAI'13)*.

Gene Ontology Consortium, T. 2000. Gene Ontology: Tool for the unification of biology. *Nature Genetics* 25:25–29.

Janowicz, K., and Wilkes, M. 2009. Sim-dla: A novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity. volume 5554, 353–367.

Lehmann, K., and Turhan, A.-Y. 2012. A framework for semantic-based similarity measures for $\mathcal{ELH}$-concepts. 307–319.

Lord, P. W.; Stevens, R. D.; Brass, A.; and Goble, C. A. 2003. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics* 19(10):1275–1283.

Lutz, C., and Wolter, F. 2010. Deciding inseparability and conservative extensions in the description logic $\mathcal{EL}$. *Journal of Symbolic Computation* 45(2):194–228.

Mistry, M., and Pavlidis, P. 2008. Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics* 9.

Peñaloza, R., and Turhan, A.-Y. 2011. A practical approach for computing generalization inferences in $\mathcal{EL}$. In Grobelnik, M., and Simperl, E., eds., *Proceedings of the 8th European Semantic Web Conference (ESWC'11)*, 410–423.

Schlicker, A.; Domingues, F. S.; Rahnenführer, J.; and Lengauer, T. 2006. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* 7:302.

Suntisrivaraporn, B. 2013. A similarity measure for the description logic $\mathcal{EL}$ with unfoldable terminologies. In *5th International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, 408–413.