Dresden University of Technology
Institute for Theoretical Computer Science
Chair for Automata Theory

LTCS–Report

Subsumption and Instance Problem in $\mathcal{EL}\mathcal{H}$ w.r.t. General TBoxes

Sebastian Brandt

LTCS-Report 04-04

Lehrstuhl für Automatentheorie
Institut für Theoretische Informatik
TU Dresden
http://lat.inf.tu-dresden.de

Hans-Grundig-Str. 25
01062 Dresden
Germany

Subsumption and Instance Problem in $\mathcal{EL}\mathcal{H}$ w.r.t. General TBoxes

Sebastian Brandt
Theoretical Computer Science
TU Dresden
brandt@tcs.inf.tu-dresden.de

Abstract

Recently, it was shown for the DL $\mathcal{EL}$ that subsumption and instance problem w.r.t. cyclic terminologies can be decided in polynomial time. In this paper, we show that both problems remain tractable even when admitting general concept inclusion axioms and simple role inclusion axioms.

Contents

1 Motivation

In the area of DL based knowledge representation, the utility of *general* TBoxes, i.e., TBoxes that allow for general concept inclusion (GCI) axioms, is well known. For instance, in the context of the medical terminology GALEN [20], GCIs are used especially for two purposes [18]:

- indicate the status of objects: instead of introducing several concepts for the same concept in different states, e.g., normal insulin secretion, abnormal but harmless insulin secretion, and pathological insulin secretion, only insulin secretion is defined while the status, i.e., normal, abnormal but harmless, and pathological is implied by GCIs of the form $\ldots \sqsubseteq \exists \mathsf{has\_status.pathological}$.

- to bridge levels of granularity and add implied meaning to concepts. A classical example [12] is to use a GCI like

$$\mathsf{ulcer} \sqcap \exists \mathsf{has\_loc.stomach} \sqsubseteq \mathsf{ulcer} \sqcap \exists \mathsf{has\_loc.}(\mathsf{lining} \sqcap \exists \mathsf{is\_part\_of.stomach})$$

  to render the description of 'ulcer of stomach' more precisely to 'ulcer of lining of stomach' if it is known that 'ulcer of stomach' is specific of the lining of the stomach.

It has been argued that the use of GCIs facilitates the re-use of data in applications of different levels of detail while retaining all inferences obtained from the full description [20]. Hence, to examine reasoning w.r.t. general TBoxes has a strong practical motivation.

Research on reasoning w.r.t. general TBoxes has mainly focused on very expressive DLs, reaching as far as, e.g., $\mathcal{ALCNR}$ [5] and $\mathcal{SHIQ}$ [13], in which deciding subsumption of concepts w.r.t. general TBoxes is EXPTIME hard. Fewer results exist on subsumption w.r.t. general terminologies DLs below $\mathcal{ALC}$. In [10] the problem is shown to remain EXPTIME complete for a DL providing only conjunction, value restriction and existential restriction. The same holds for the small DL $\mathcal{AL}$ which allows for conjunction, value and unqualified existential restriction, and primitive negation [8]. Even for the simple DL $\mathcal{FL}_0$, which only allows for conjunction and value restriction, subsumption w.r.t. cyclic TBoxes with descriptive semantics is PSPACE hard [15], implying hardness for general TBoxes.

Recently, however, it was shown for the DL $\mathcal{EL}$ that subsumption and instance problem w.r.t. cyclic terminologies can be decided in polynomial time [4, 3]. In the present paper we show that even w.r.t. general $\mathcal{ELH}$-TBoxes, including GCIs and simple role inclusion axioms, subsumption and instance problem remain tractable. A surprising result given that DL systems usually employed for reasoning over general terminologies implement—highly optimized—EXPTIME algorithms [14, 11]. Similarly, RACER [11], the only practicable reasoner for ABox reasoning w.r.t. general TBoxes uses an EXPTIME algorithm for the very expressive DL $\mathcal{ALCNH}_{R^+}$.

The paper is organized as follows. Basic definitions related to general $\mathcal{ELH}$ TBoxes are introduced in Section 2. In Sections 3 and 4 we show how to decide subsumption and instance problem, respectively, w.r.t. general $\mathcal{ELH}$-TBoxes in polynomial time.

2 General TBoxes in $\mathcal{ELH}$

Concept descriptions are inductively defined with the help of a set of concept *constructors*, starting with a set N_{con} of *concept names* and a set N_{role} of *role names*. In this paper, we consider the DL $\mathcal{ELH}$ which provides the concept constructors top-concept ($\top$), conjunction ($C \sqcap D$), and existential restrictions ($\exists r.C$).

As usual, the semantics of concept descriptions is defined in terms of an *interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$. The domain $\Delta^{\mathcal{I}}$ of $\mathcal{I}$ is a non-empty set and the interpretation function $\cdot^{\mathcal{I}}$ maps each concept name $P \in N_{\text{con}}$ to a subset $P^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and each role name $r \in N_{\text{role}}$ to a binary relation $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. The extension of $\cdot^{\mathcal{I}}$ to arbitrary concept descriptions is defined inductively as follows.

$$\begin{aligned} \top^{\mathcal{I}} &:= \Delta^{\mathcal{I}} \\ (C \sqcap D)^{\mathcal{I}} &:= C^{\mathcal{I}} \cap D^{\mathcal{I}} \\ (\exists r.C)^{\mathcal{I}} &:= \{x \in \Delta^{\mathcal{I}} \mid \exists y\colon (x,y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\} \end{aligned}$$

For a given the DL $\mathcal{L}$, an $\mathcal{L}$-terminology (called $\mathcal{L}$-TBox) is a finite set $\mathcal{T}$ of axioms of the form $C \sqsubseteq D$ (called *GCI*) or $C \doteq D$ (called *definition*) or $r \sqsubseteq s$ (called *simple role inclusion axiom* (SRI)), where C and D are concept descriptions defined in $\mathcal{L}$ and $r, s \in N_{\text{role}}$. A concept name $A \in N_{\text{con}}$ is called *defined in* $\mathcal{T}$ iff $\mathcal{T}$ contains one or more axioms of the form $A \sqsubseteq D$ or $A \doteq D$.

The *size* of $\mathcal{T}$ is defined as the sum of the sizes of all axioms in $\mathcal{T}$. Denote by $N_{\text{con}}^{\mathcal{T}}$ the set of all concept names occurring in $\mathcal{T}$ and by $N_{\text{role}}^{\mathcal{T}}$ the set of all role names occurring in $\mathcal{T}$. A TBox that contains GCIs is called *general*. Denote by $\mathcal{ELH}$ the extension of $\mathcal{EL}$ by SRIs in TBoxes.

An interpretation $\mathcal{I}$ is a *model* of $\mathcal{T}$ iff for every GCI $C \sqsubseteq D \in \mathcal{T}$ it holds that $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$, for every definition $C \doteq D$ it holds that $C^{\mathcal{I}} = D^{\mathcal{I}}$, and for every SRI $r \sqsubseteq s$ it holds that $r^{\mathcal{I}} \subseteq s^{\mathcal{I}}$. A concept description C is *satisfiable* w.r.t. $\mathcal{T}$ iff there exists a model $\mathcal{I}$ such that $C^{\mathcal{I}} \neq \emptyset$. A concept description C *subsumes* a concept description D w.r.t. $\mathcal{T}$ ($C \sqsubseteq_{\mathcal{T}} D$) iff $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ in every model $\mathcal{I}$ of $\mathcal{T}$. C and D are *equivalent* w.r.t. $\mathcal{T}$ ($C \equiv_{\mathcal{T}} D$) iff they subsume each other w.r.t. $\mathcal{T}$.

An $\mathcal{L}$-ABox is a finite set of assertions of the form $A(a)$ (called *concept assertion*) or $r(a, b)$ (called *role assertion*), where $A \in N_{\text{con}}$, $r \in N_{\text{role}}$, and a, b are *individual names* from a set N_{ind}. $\mathcal{I}$ is a model of a TBox $\mathcal{T}$ together with an ABox $\mathcal{A}$ iff $\mathcal{I}$ is a model of $\mathcal{T}$ and $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ such that all assertions in $\mathcal{A}$ are satisfied, i.e., $a^{\mathcal{I}} \in A^{\mathcal{I}}$ for all $A(a) \in \mathcal{A}$ and $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in r^{\mathcal{I}}$ for all $r(a, b) \in \mathcal{A}$. An individual name a *is an instance of* C w.r.t. $\mathcal{T}$ ($\mathcal{A} \models_{\mathcal{T}} C(a)$) iff $a^{\mathcal{I}} \in A^{\mathcal{I}}$ for all models $\mathcal{I}$ of $\mathcal{T}$ together with $\mathcal{A}$. Denote by $N_{\text{ind}}^{\mathcal{A}}$ the set of all individual names occurring in an ABox $\mathcal{A}$.

The above semantics for TBoxes and ABoxes is usually called *descriptive semantics* [17]. In case of an empty TBox, we write $C \sqsubseteq D$ instead of $C \sqsubseteq_{\emptyset} D$ and analogously $C \equiv D$ instead of $C \equiv_{\emptyset} D$.

Example 1 As an example of what can be expressed with an $\mathcal{ELH}$-TBox, consider the following TBox showing in an extremely simplified fashion a part of a medical terminology.

$$\begin{aligned}
\mathsf{Pericardium} &\sqsubseteq \mathsf{Tissue} \sqcap \exists \mathsf{cont_in}.\mathsf{Heart} \\
\mathsf{Pericarditis} &\sqsubseteq \mathsf{Inflammation} \\
&\quad \sqcap \exists \mathsf{has_loc}.\mathsf{Pericardium} \\
\mathsf{Inflammation} &\sqsubseteq \mathsf{Disease} \sqcap \exists \mathsf{acts_on}.\mathsf{Tissue} \\
\mathsf{Disease} \sqcap \exists \mathsf{has_loc}.\exists \mathsf{comp_of}.\mathsf{Heart} &\sqsubseteq \mathsf{Heartdisease} \\
&\quad \sqcap \exists \mathsf{is_state}.\mathsf{NeedsTreatment} \\
\mathsf{cont_in} &\sqsubseteq \mathsf{comp_of}
\end{aligned}$$

The TBox contains four GCIs and one SRI, stating, e.g., that Pericardium is tissue contained in the heart and that a diesease located in a component

of the heart is a heart disease and requires treatment. Without going into detail, one can check that Pericarditis would be classified as a heart disease requiring treatment because, as stated in the TBox, Pericarditis is a disease located in the Pericardium contained in the heart, and everything contained in something is a component of it.[1]

3 Subsumption in $\mathcal{ELH}$ with GCIs

We aim to show that subsumption of $\mathcal{ELH}$ concepts w.r.t. general TBoxes can be decided in polynomial time. A natural question is whether we may not simply utilize an existing decision procedure for a more expressive DL which might exhibit polynomial time complexity when applied to $\mathcal{ELH}$ TBoxes. Using the standard tableaux algorithm deciding consistency of general $\mathcal{ALC}$-TBoxes [2] as an example, one can show that this approach in general does not bear fruit, eveb for the sublanguage $\mathcal{EL}$.

In order to decide subsumption $C \sqsubseteq^?_{\mathcal{T}} D$ w.r.t. an $\mathcal{EL}$-TBox, an intuitive decision procedure to choose would be the $\mathcal{ALC}$ tableaux algorithm deciding consistency of $\mathcal{ALC}$-concepts w.r.t. $\mathcal{ALC}$ terminologies [1]. The DL $\mathcal{ALC}$ extends $\mathcal{EL}$ by value restrictions ($\forall$), disjunction ($\sqcup$), and negation ($\neg$). We can decide $C \sqsubseteq^?_{\mathcal{T}} D$ by deciding satisfiability of $C \sqcap \neg D$ w.r.t. $\mathcal{T}$.

The following example presents a general $\mathcal{EL}$-TBox for which the $\mathcal{ALC}$ tableaux algorithm takes exponentially many steps in the worst case. We use the standard $\mathcal{ALC}$ tableaux as describied in [1].

Example 2 For $n \in \mathbb{N}$, let $N_{\text{con}} := \{A, B, C, D\} \cup \{A_i \mid 1 \leq i \leq n\} \cup \{B_i \mid 1 \leq i \leq n\}$ and $N_{\text{role}} := \{r\}$. Define the TBox $\mathcal{T}_n$ as follows:

$$
\begin{aligned}
C &\doteq A \\
D &\doteq \exists r.B \\
\exists r.B &\sqsubseteq B \\
A &\sqsubseteq \exists r.A \\
\exists r.A_i \sqcap \exists r.B_i &\sqsubseteq B \qquad \text{for every} 1 \leq i \leq n
\end{aligned}
$$

[1] The example is only supposed to show the features of $\mathcal{ELH}$ and in no way claims to be adequate from a Medical KR point of view.

To be able to apply the tableaux algorithm, the GCIs in $\mathcal{T}_n$ are represented as tautologies:

$$\begin{aligned} B \sqcup \forall r.\neg B & \\ \neg A \sqcup \exists r.A & \\ B \sqcup \forall r.\neg A_i \sqcup \forall r.\neg B_i & \qquad \text{for every} 1 \leq i \leq n \end{aligned}$$

Figure 1 shows (in an abridged way) the first four steps of the tableaux computation for $\mathcal{T}$. The tableaux algorithm starts in Step 0 with a model of one vertex x_0 labeled by $C \sqcap \neg D$. A so-called 'blocking' techniqe is used to avoid the generation of infinitely many vertices for a model: if the label of the new vertex w is a subset of a label of an old vertex v then w is removed, redirecting the edge pointing to w to the old vertex v.

Since x_0 could not be blocked, all GCIs are added to the label of x_0, yielding the situation denoted as Step 1 in Figure 1. In the tableaux, disjunction is dealt with by means of nondeterminism: a GCI of the form $C \sqcup D$ is resolved by nondeterministically choosing between C or D to add to the label set of the vertex under consideration (see [1] for details). Since the concept name A is already contained in the label of x_0, the only possibility to satisfy the GCI $\neg A \sqcup \exists r.A$ (shown boxed in Step 1) is to introduce an r-successor x_1 to x_0. Several other GCIs in the label of x_0 have to be satisfied. In particular, if the algorithm chooses the disjunct $\forall r.\neg B$ from the GCI $B \sqcup \forall r.\neg B$ then $\neg B$ is added to the label set of x_1. Moreover, for every $1 \leq i \leq n$ the GCI

$$B \sqcup \forall r.\neg A_i \sqcup \forall r.\neg B_i \qquad 1 \leq i \leq n$$

must be satisfied for x_0. Since $\neg B$ is already in the label of x_1, thus ruling out choosing B, the algorithm for every i has to include either $\neg A_i$ or $\neg B_i$ into the label of x_1. Hence a set S_1 is added to the label set of x_1, where S_1 corresponds to a tuple $\bar{s}_1$ with

$$\bar{s}_1 \in \{\neg A_1, \neg B_1\} \times \cdots \times \{\neg A_n, \neg B_n\} =: S. \qquad (*)$$

Without going into detail further, Steps 3 and 4 in Figure 1 illustrate that the tableaux algorithm necessarily adds a successor x_2 of x_1 whose label set consists of $A, \neg B$ and another set S_2 representing another nondeterministic choice from S, see $(*)$. Hence, the introduction of x_2 can be blocked only if the algorithm nondeterministically chose $S_1 = S_2$.

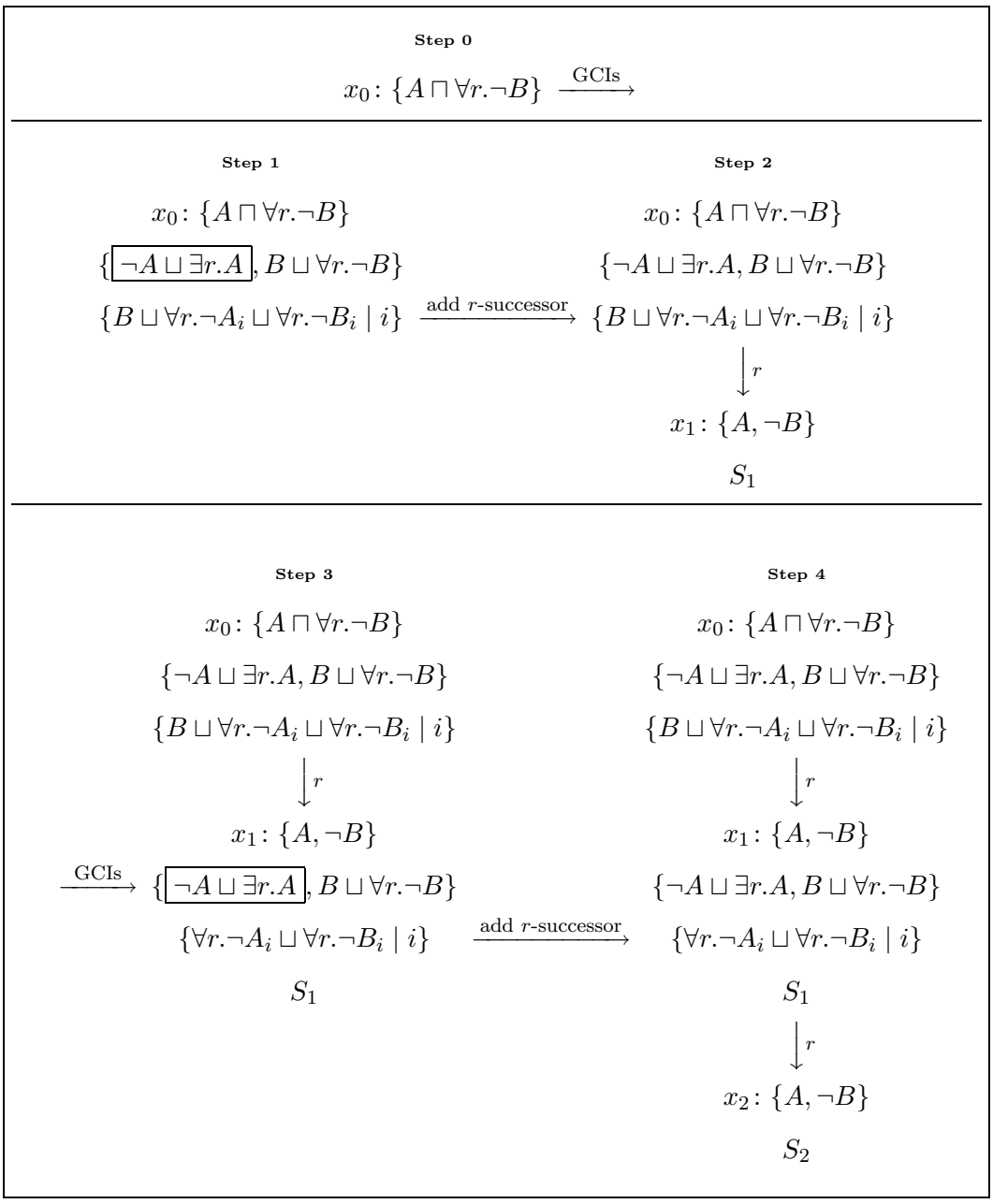

Figure 1: $\mathcal{ALCI}$ tableaux computation

Obviously, the situation for x_2 resembles that of x_1, implying that another successor x_3 is introduced and so on. As there exist exponentially many sets S_j mutually incomparable w.r.t. the subset relation the nondeterminism of the tableaux algorithm might give rise to an exponentially long line of successors before a vertex x_k is introduced in whose label the set S_k *necessarily* is a repitition of a label set seen before.

Hence, the standard tableaux algorithm in the worst case needs exponentially many steps to decide the subsumption $C \sqsubseteq_{\mathcal{T}} D$.

Hence, new techniques are required exploiting the simpler structure of general $\mathcal{ELH}$-TBoxes better. The first step in our approach is to transform TBoxes into a normal form which limits the use of complex concept descriptions to the most basic cases.

Definition 3 (Normalized $\mathcal{ELH}$ TBox) Let $\mathcal{T}$ be an $\mathcal{ELH}$-TBox over N_{con} and N_{role}. $\mathcal{T}$ is *normalised* iff (i) $\mathcal{T}$ contains only GCIs and SRIs, and, (ii) all of the GCIs have one of the following forms:

$$\begin{aligned} A &\sqsubseteq B \\ A_1 \sqcap A_2 &\sqsubseteq B \\ A &\sqsubseteq \exists r.B \\ \exists r.A &\sqsubseteq B. \end{aligned}$$

where A, A_1, A_2, B represent concept names from $N_{\text{con}}^{\top}$.

Such a normal form can easily be computed in polynomial time and does not increase the size of the TBox more than polynomially. The following definition provides normalization rules by which an arbitrary $\mathcal{EL}$-TBox can be transformed into a normalized one. The normalization rules are inspired by [16] where a similar problem is solved for $\mathcal{ALC}$-TBoxes containing only definitions.

Definition 4 (Normalization rules) Let $\mathcal{T}$ be an $\mathcal{ELH}$-TBox over N_{con} and N_{role}. For every $\mathcal{ELH}$-concept description C, D, E over $N_{\text{role}} \cup \{\top\}$ and for every $r \in N_{\text{role}}$, the $\mathcal{ELH}$-normalization rules are defined modulo commutativity

of conjunction ($\sqcap$) as follows:

$$\begin{array}{lrcl}
\textsf{NF1} & C \doteq D & \longrightarrow & \{C \sqsubseteq D,\, D \sqsubseteq C\} \\
\textsf{NF2} & \hat{C} \sqcap D \sqsubseteq E & \longrightarrow & \{\hat{C} \sqsubseteq A,\, A \sqcap D \sqsubseteq E\} \\
\textsf{NF3} & \exists r.\hat{C} \sqsubseteq D & \longrightarrow & \{\hat{C} \sqsubseteq A,\, \exists r.A \sqsubseteq D\} \\
\textsf{NF4} & C \sqsubseteq \exists r.\hat{D} & \longrightarrow & \{C \sqsubseteq \exists r.A,\, A \sqsubseteq \hat{D}\} \\
\textsf{NF5} & C \sqsubseteq D \sqcap E & \longrightarrow & \{C \sqsubseteq D,\, C \sqsubseteq E\}
\end{array}$$

where $\hat{C}, \hat{D}$ denote non-atomic concept descriptions and A denotes a new concept name from N_{con}. Applying a rule $G \longrightarrow \mathcal{S}$ to $\mathcal{T}$ changes $\mathcal{T}$ to $(\mathcal{T} \setminus \{G\}) \cup \mathcal{S}$. The normalized TBox $norm(\mathcal{T})$ is defined by exhaustively applying Rules **NF1** to **NF3** and, after that, exhaustively applying Rules **NF4** and **NF5**.

The size of $\mathcal{T}$ is increased only linearly by exhaustive application of Rule **NF1**. Since this rule never becomes applicable as a consequence of Rules **NF2** to **NF5**, we may restrict our attention to Rules **NF2** to **NF5**. A single application of one of the Rules **NF2** to **NF3** increases the size of $\mathcal{T}$ only by a constant, introducing a new concept name and splitting one GCI into two. Exhaustive application therefore produces an ontology of linear size in the size of $\mathcal{T}$.

After exhaustive application of Rules **NF1** to **NF3**, the left-hand side of every GCI is of constant size. Hence, applying Rules **NF4** and **NF5** exhaustively similarly yields an ontology of linear size in $\mathcal{T}$. Conseqently, the following lemma holds.

Lemma 5 *The normalized TBox $norm(\mathcal{T})$ can be computed in linear time in the size of $\mathcal{T}$. The resulting ontology is of linear size in the size of $\mathcal{T}$.*

Our strategy is now, for every concept name $A \in N_{\text{con}}^{\mathcal{T}}$ and $\top$, to compute a set of concept names $S_*(A)$ with the following property: whenever in some point x in a model of $\mathcal{T}$ the concept A holds then every concept in $S_*(A)$ necessarily also holds in x. Similarly, for every role r we want to represent by $S_*(r)$ the set of all roles included in r. The simple structure of GCIs in normalized TBoxes allows us to define such sets as follows. To simplify Notation, let $N_{\text{con}}^{\mathcal{T},\top} := N_{\text{con}}^{\mathcal{T}} \cup \{\top\}$.

ISR If $s \in S_i(r)$ and $s \sqsubseteq t \in \mathcal{T}$ and $t \notin S_{i+1}(r)$
then $S_{i+1}(r) := S_{i+1}(r) \cup \{t\}$

IS1 If $A_1 \in S_i(A)$ and $A_1 \sqsubseteq B \in \mathcal{T}$ and $B \notin S_{i+1}(A)$
then $S_{i+1}(A) := S_{i+1}(A) \cup \{B\}$

IS2 If $A_1, A_2 \in S_i(A)$ and $A_1 \sqcap A_2 \sqsubseteq B \in \mathcal{T}$
and $B \notin S_{i+1}(A)$ then $S_{i+1}(A) := S_{i+1}(A) \cup \{B\}$

IS3 If $A_1 \in S_i(A)$ and $A_1 \sqsubseteq \exists r.B \in \mathcal{T}$
and $B_1 \in S_i(B)$ and $s \in S_i(r)$ and $\exists s.B_1 \sqsubseteq C \in \mathcal{T}$
and $C \notin S_{i+1}(A)$ then $S_{i+1}(A) := S_{i+1}(A) \cup \{C\}$

Figure 2: Rules for implication sets

Definition 6 (Implication set) Let $\mathcal{T}$ denote a normalized $\mathcal{ELH}$-TBox $\mathcal{T}$ over N_{con} and N_{role}. For every $A \in N_{\text{con}}^{\mathcal{T},\top}$ ($r \in N_{\text{role}}^{\mathcal{T}}$) and every $i \in \mathbb{N}$, the set $S_i(A)$ ($S_i(r)$) is defined inductively, starting by $S_0(A) := \{A, \top\}$ ($S_0(r) := \{r\}$). For every $i \geq 0$, $S_{i+1}(A)$ ($S_{i+1}(r)$) is obtained by extending $S_i(A)$ ($S_i(r)$) by exhaustive application of the extension rules shown in Figure 2. The *implication set* $S_*(A)$ of A is defined as the infinite union $S_*(A) := \bigcup_{i \geq 0} S_i(A)$. Analogously, $S_*(r) := \bigcup_{i \geq 0} S_i(r)$.

Note that the successor $S_{i+1}(A)$ of some $S_i(A)$ is generally not the result of only a *single* rule application. $S_{i+1}(A)$ is complete only if no more rules are applicable to any $S_i(B)$ or $S_i(r)$. Implication sets induce a reflexive and transitive but not symmetric relation on $N_{\text{con}}^{\mathcal{T},\top}$ and $N_{\text{role}}^{\mathcal{T}}$, since $B \in S_*(A)$ does not imply $A \in S_*(B)$.

We have to show that the idea underlying implication sets is indeed correct. Hence, the occurrence of a concept name B in $S_*(A)$ implies that $A \sqsubseteq_{\mathcal{T}} B$ and vice versa.

Theorem 7 *For every normalised $\mathcal{ELH}$-TBox over N_{con} and N_{role}, (i) for every $r, s \in N_{\text{role}}^{\mathcal{T}}$, $s \in S_*(r)$ implies $r \sqsubseteq_{\mathcal{T}} s$, and (ii) for every $A, B \in N_{\text{con}}^{\mathcal{T},\top}$ it holds that $B \in S_*(A)$ iff $A \sqsubseteq_{\mathcal{T}} B$.*

Proof. (i) Proof by induction over n. As $S_0(r) = \{r\}$, the claim holds trivially. For $n > 0$ we know by Rule **ISR** that there exists a role $t \in S_{n-1}(r)$ and a SRI $t \sqsubseteq s \in \mathcal{T}$. By induction hypothesis $r \sqsubseteq_{\mathcal{T}} t$ which by transitivity

of role inclusion axioms yields $r \sqsubseteq_{\mathcal{T}} s$. For the reverse direction, $r \sqsubseteq_{\mathcal{T}} s$ immediately implies a finite chain

$$\{r \sqsubseteq t_0\} \cup \{t_i \sqsubseteq t_{i+1} \mid 0 \leq i \leq k-1\} \cup \{t_k \sqsubseteq s\} \subseteq \mathcal{T}$$

of SRIs in $\mathcal{T}$, implying by a finite number of applications of Rule **ISR** that $s \in S_{k+1}(r)$.

(ii) ($\Rightarrow$) It suffices to show for every model $\mathcal{I}$ of $\mathcal{T}$ and for every $B \in S_*(A)$ that $x \in A^{\mathcal{I}}$ implies $x \in B^{\mathcal{I}}$. Assume a model $\mathcal{I}$ of $\mathcal{T}$ with a witness $x \in A^{\mathcal{I}}$ and let $B \in S_*(A)$. Proof by induction over n where n is the least index with $B \in S_n(A)$.

($n = 0$) Then, $S_n(A) = \{A\}$ implying $B = A$. As x was chosen a witness of A the claim holds.

($n > 0$) In Step $n-1$, B can have been included into $S_n(A)$ by any of the Rules **IS1** to **IS6**. We distinguish one case for each rule.

(**IS1**) There exists a concept name $A_1 \in S_{n-1}(A)$ and a GCI $G := A_1 \sqsubseteq B \in \mathcal{T}$. By induction hypothesis (IH), $x \in A_1^{\mathcal{I}}$, implying by G that also $x \in B^{\mathcal{I}}$.

(**IS2**) There exist two concept names $A_1, A_2 \in S_{n-1}(A)$ and a GCI $G := A_1 \sqcap A_2 \sqsubseteq B \in \mathcal{T}$. By IH, $A_1, A_2 \in S_{n-1}(A)$ yields $x \in A_1^{\mathcal{I}}$ and $x \in A_2^{\mathcal{I}}$, implying by G that $x \in B^{\mathcal{I}}$.

(**IS3**) There exist concept names $A_1 \in S_{n-1}(A)$, $A_2 \in N_{\text{con}}^{\mathcal{T},\top}$, and $A_3 \in S_{n-1}(A_2)$ and two GCIs $G := A_1 \sqsubseteq \exists r.A_2$ and $H := \exists s.A_3 \sqsubseteq B$ with $s \in S_{n-1}(r)$. By IH, $r \sqsubseteq s$, implying by G that $x \in (\exists r.A_2)^{\mathcal{I}}$. Since $A_3 \in S_{n-1}(A_2)$ and , the IH implies $x \in A_1^{\mathcal{I}}$ and $x \in (\exists s.A_3)^I$, yielding by H that $x \in B^{\mathcal{I}}$.

($\Leftarrow$) It suffices to show that if $B \notin S_*(A)$ then we can construct a model $\mathcal{I}$ of $\mathcal{T}$ with a witness $x \in A^{\mathcal{I}} \setminus B^{\mathcal{I}}$.

We construct a (possibly infinite) *canonical model* $\mathcal{I}(A)$ of A w.r.t. $\mathcal{T}$ by means of the following definition. $I(A)$ is defined iteratively starting by $I_0(A)$. Define $\Delta^{\mathcal{I}_0(A)} := \{x_A\}$ and $B^{\mathcal{I}_0(A)} := \{x_A \mid B = A\}$ for all $B \in N_{\text{con}}^{\mathcal{T},\top}$. For $i \geq 0$, the model $\mathcal{I}_{i+1}$ is defined as an extension of $\mathcal{I}_i$ obtained by exhaustive application of the following generation rules.

CM1 If $A \sqsubseteq B \in \mathcal{T}$ then, for every individual $x \in \Delta^{\mathcal{I}_i}$ with $x \in A^{\mathcal{I}_i}$ and $x \notin B^{\mathcal{I}_i}$, add x to $B^{\mathcal{I}_{i+1}}$

CM2 If $A \sqcap B \sqsubseteq C \in \mathcal{T}$ then, for every individual $x \in \Delta^{\mathcal{I}_i}$ with $x \in A^{\mathcal{I}_i} \cap B^{\mathcal{I}_i}$ and $x \notin C^{\mathcal{I}_{i+1}}$, add x to $C^{\mathcal{I}_{i+1}}$

CM3 If $A \sqsubseteq \exists r.B \in \mathcal{T}$ then, for every individual $x \in \Delta^{\mathcal{I}_i}$ with $x \in A^{\mathcal{I}_i}$ for which no r-successor $y \in \Delta^{\mathcal{I}_{i+1}}$ with $y \in B^{\mathcal{I}_{i+1}}$ exists, introduce a new individual y to $\Delta^{\mathcal{I}_{i+1}}$ and include y into $B^{\mathcal{I}_{i+1}}$ and include (x, y) into $r^{\mathcal{I}_{i+1}}$

CM4 If $\exists r.A \sqsubseteq B \in \mathcal{T}$ then, for every pair $(x, y) \in s^{\mathcal{I}_i}$ with $s \sqsubseteq_{\mathcal{T}} r$ and $y \in A^{\mathcal{I}_i}$ and $x \notin B^{\mathcal{I}_{i+1}}$, include x into $B^{\mathcal{I}_{i+1}}$

The above rules are applied fairly, i.e., every rule applicable to already existing elements $x \in \Delta^{\mathcal{I}_i}$ will be applied before applying rules to new elements. The canonical model $\mathcal{I}(A)$ is defined as the infinite union $\mathcal{I}(A) := \bigcup_{i \geq 0} \mathcal{I}_i(A)$.

We first prove that $\mathcal{I}(A)$ in fact is a model of A w.r.t. $\mathcal{T}$. Assume that $x_A \notin A^{\mathcal{I}(A)}$. In this case there is a $y \in \Delta^{\mathcal{I}(A)}$ for which a GCI $G \in \mathcal{T}$ is violated. As $\mathcal{T}$ is normalized, it suffices to distinguish four cases for the violated GCI G.

- If $G = B \sqsubseteq C \in \mathcal{T}$ then $y \in B^{\mathcal{I}(A)}$ but $y \notin C^{\mathcal{I}(A)}$. Consider the least index n with $y \in B^{\mathcal{I}_n(A)}$. By definition, Rule **CM1** causes y to be added to $C^{\mathcal{I}_{n+1}} \subseteq C^{\mathcal{I}}$, contradicting the assumption.

- If $G = B \sqcap C \sqsubseteq D \in \mathcal{T}$ then $y \in B^{\mathcal{I}(A)} \sqcap C^{\mathcal{I}(A)}$ but $y \notin D^{\mathcal{I}(A)}$. Consider the least index n with $y \in B^{\mathcal{I}_n(A)} \sqcap C^{\mathcal{I}_n(A)}$. Rule **CM2** causes y to be added to $D^{\mathcal{I}_{n+1}} \subseteq D^{\mathcal{I}}$, in contradiction to the assumption.

- If $G = B \sqsubseteq \exists r.C \in \mathcal{T}$ then $y \in B^{\mathcal{I}(A)}$ but y has no appropriate r-successor. Consider the least n with $y \in B^{\mathcal{I}_n(A)}$. By Rule **CM3**, a new element z is introduced to $\Delta^{\mathcal{I}_{n+1}}$, the pair (y, z) added to $r^{\mathcal{I}_{n+1}}$, and z added to $C^{\mathcal{I}_{n+1}}$, again in contradiction to the assumption.

- If $G = \exists r.B \sqsubseteq C \in \mathcal{T}$ then there exists an edge $(y, z) \in s^{\mathcal{I}(A)}$ with $s \sqsubseteq_{\mathcal{T}} r$ such that $z \in B^{\mathcal{I}(A)}$ but $y \notin C^{\mathcal{I}(A)}$. Consider the least n with $z \in B^{\mathcal{I}_n(A)}$. As $s \sqsubseteq_{\mathcal{T}} r$ and $(y, z) \in s^{\mathcal{I}_n(A)}$, Rule **CM4** adds y to $C^{\mathcal{I}_{n+1}(A)} \subseteq C^{\mathcal{I}(A)}$, contradicting the assumption.

Having proven $\mathcal{I}(A)$ to be a model of A w.r.t. $\mathcal{T}$ it remains to show that $B^{\mathcal{I}(A)} \not\subseteq A^{\mathcal{I}(A)}$. To this end, we show for every $n \in \mathbb{N}$, for *every* $A, B \in N_{\text{con}}^{\mathcal{T},\top}$, $A \neq B$, and for *every* $x \in A^{\mathcal{I}_n(A)}$: if $\{C \mid C \in x^{\mathcal{I}_t(A)}\} = \{A\}$ for some minimally chosen $t \in \mathbb{N}$ and $x \in B^{\mathcal{I}_n(A)}$ then $B \in S_*(A)$. Note that $B \in S_*(A)$ holds if $B \in S_m(A)$ for some $m \in \mathbb{N}$ since $S_m(A) \subseteq S_*(A)$.

($n = 0$) Trivial since $B^{\mathcal{I}_0(A)} = \emptyset$ implies that the premise $x \in B^{\mathcal{I}_n(A)}$ does not hold.

($n \geq 0$) Let $\{C \mid C \in x^{\mathcal{I}_t(A)}\} = \{A\}$ for some $t < n$ and let $x \in B^{\mathcal{I}_n(A)} \setminus B^{\mathcal{I}_{n-1}(A)}$. In the definition of $\mathcal{I}_n(A)$ there are four rules which can have caused the inclusion of x into $B^{\mathcal{I}_n(A)}$:

- (**CM1**) Then then there is a GCI $G := A_1 \sqsubseteq B \in \mathcal{T}$ and $x \in A_1^{\mathcal{I}_{n-1}(A)}$. If $t = n-1$ then $A_1 = A$, implying $B \in S_1(A)$ by Rule **IS1** with G. If $t < n-1$ then, by induction hypothesis (IH), $A_1 \in S_*(A)$, implying $A_1 \in S_m(A)$ for some $m \in \mathbb{N}$, yielding $B \in S_{m+1}(A)$ by Rule **IS1** with G.

- (**CM2**) Then there is a GCI $G := A_1 \sqcap A_2 \sqsubseteq B \in \mathcal{T}$ and $x \in A_1^{\mathcal{I}_{n-1}(A)} \cap A_2^{\mathcal{I}_{n-1}(A)}$. If $t = n-1$ then $A_1 = A_2 = A$, implying $B \in S_1(A)$ by Rule **IS2** with G. If $t < n-1$ then, by IH, $\{A_1, A_2\} \subseteq S_*(A)$. Hence, $\{A_1, A_2\} \subseteq S_m(A)$ for some $m \in \mathbb{N}$, implying $B \in S_{m+1}(A)$ by Rule **IS2** with G.

- (**CM4**) Then there is a GCI $G := \exists r.A_1 \sqsubseteq B \in \mathcal{T}$ and $y \in \Delta^{\mathcal{I}_{n-1}(A)}$ with $(x, y) \in s^{\mathcal{I}_{n-1}(A)}$ with $s \sqsubseteq_{\mathcal{T}} r$ and $y \in A_1^{\mathcal{I}_{n-1}(A)}$, implying $t < n-1$ since x and y cannot be created at the same time. Hence, firstly, there is a GCI $H := C \sqsubseteq \exists s.D \in \mathcal{T}$ and an index $t \leq k < n-1$ with $x \in C^{\mathcal{I}_k}$, implying $(x, y) \in s^{\mathcal{I}_{k+1}(A)}$ and $y \in D^{\mathcal{I}_{k+1}}$. Secondly, $y \in A_1^{\mathcal{I}_{k+1}}$. By IH, $A_1 \in S_*(D)$. If $t = k$ then $C = A$, otherwise, by IH, $C \in S_*(A)$. In both cases there exists a least index m with $C \in S_m(A)$ and $A_1 \in S_m(D)$, implying $B \in S_{m+1}(A)$ by Rule **IS3** with G and H.

∎

We have shown how to decide subsumption w.r.t. general $\mathcal{ELH}$-TBoxes. It remains to show that our decision procedure works in polynomial time. In contrast to the correctness proof this is relatively easy.

Lemma 8 *For every normalised $\mathcal{ELH}$-TBox over N_{con} and N_{role} and for every $A \in N_{\text{con}}^{\mathcal{T},\top}$, the implication set $S_*(A)$ can be computed in polynomial time in the size of $\mathcal{T}$.*

Proof. To show decidability in polynomial time it suffices to show that, (i) $\mathcal{T}$ can be normalized in polynomial time (see above), and, (ii) for all $A \in N_{\text{con}}^{\mathcal{T},\top}$ and $r \in N_{\text{role}}^{\mathcal{T}}$, the sets $S_*(A)$ and $S_*(r)$ can be computed in polynomial time in the size of $\mathcal{T}$. Every $S_{i+1}(A)$ and $S_{i+1}(r)$ depends only on sets with index i. Hence, once $S_{i+1}(A) = S_i(A)$ and $S_{i+1}(r) = S_i(r)$

ISR If $s \in S_i(r)$ and $s \sqsubseteq t \in \mathcal{T}$ and $t \notin S_{i+1}(r)$
then $S_{i+1}(r) := S_{i+1}(r) \cup \{t\}$

IS1 If $A_1 \in S_i(\alpha)$ and $A_1 \sqsubseteq B \in \mathcal{T}$ and $B \notin S_{i+1}(\alpha)$
then $S_{i+1}(\alpha) := S_{i+1}(\alpha) \cup \{B\}$

IS2 If $A_1, A_2 \in S_i(\alpha)$ and $A_1 \sqcap A_2 \sqsubseteq B \in \mathcal{T}$
and $B \notin S_{i+1}(\alpha)$ then $S_{i+1}(\alpha) := S_{i+1}(\alpha) \cup \{B\}$

IS3 If $A_1 \in S_i(\alpha)$ and $A_1 \sqsubseteq \exists r.B \in \mathcal{T}$
and $B_1 \in S_i(B)$ and $s \in S_i(r)$ and $\exists s.B_1 \sqsubseteq C \in \mathcal{T}$
and $C \notin S_{i+1}(\alpha)$ then $S_{i+1}(\alpha) := S_{i+1}(\alpha) \cup \{C\}$

IS4 If $r(a, b) \in \mathcal{A}$ and $B \in S_i(b)$ and $s \in S_i(r)$
and $\exists s.B \sqsubseteq C \in \mathcal{T}$ and $C \notin S_{i+1}(a)$
then $S_{i+1}(a) := S_{i+1}(a) \cup \{C\}$

Figure 3: Rules for implication sets (subsumption and instance problem)

holds for all A, r the complete implication sets are obtained. This happens after a polynomial number of steps, since $S_i(A) \subseteq N_{\text{con}}^{\mathcal{T}}$ and $S_i(r) \subseteq N_{\text{role}}^{\mathcal{T}}$. To compute $S_{i+1}(A)$ and $S_{i+1}(r)$ from the $S_i(B)$ and $S_i(s)$ costs only polynomial time in the size of $\mathcal{T}$. ∎

Theorem 9 *Subsumption in $\mathcal{ELH}$ w.r.t. GCIs can be decided in polynomial time.*

4 The instance problem in $\mathcal{ELH}$ with GCIs

We show that the instance problem in $\mathcal{ELH}$ w.r.t. general TBoxes can be decided in polynomial time. To this end, the approach to decide subsumption by means of implication sets for concept names presented in the previous section is extended to ABox individuals. For every individual name $a \in N_{\text{ind}}^{\mathcal{A}}$, we want to compute a set $S_*(a)$ of concept names with the following property: if $A \in S_*(a)$ then in every model $\mathcal{I}$ of $\mathcal{T}$ together with $\mathcal{A}$ the individual $a^{\mathcal{I}}$ is a witness of A (and vice versa). To extend the definition of implication sets in this way we generalize Rules **IS1** to **IS3** to individual names and introduce a new Rule **IS4** specifically for individual names.

Definition 10 (Implication set) Let $\mathcal{T}$ denote a normalized $\mathcal{ELH}$-TBox $\mathcal{T}$ over N_{con} and N_{role} and $\mathcal{A}$ an ABox over N_{ind}, $N_{\text{con}}^{\mathcal{T}}$ and $N_{\text{role}}^{\mathcal{T}}$. For every

$A \in N_{\text{con}}^{\mathcal{T},\top}$, $r \in N_{\text{role}}^{\mathcal{T}}$, and $a \in N_{\text{ind}}^{\mathcal{A}}$ and every $i \in \mathbb{N}$, the sets $S_i(A)$, $S_i(r)$, and $S_i(a)$ are defined inductively, starting by $S_0(A) := \{A, \top\}$, $S_0(r) := \{r\}$, and $S_0(a) := \{A \mid A(a) \in \mathcal{A}\} \cup \{\top\}$, respectively. For every $i \geq 0$, $S_{i+1}(A)$, $S_{i+1}(r)$, and $S_{i+1}(a)$ are obtained by extending $S_i(A)$, $S_i(r)$, and $S_i(a)$, respectively, by exhaustive application of the extension rules shown in Figure 3, where $\alpha \in N_{\text{con}}^{\mathcal{T},\top} \cup N_{\text{ind}}^{\mathcal{A}}$. The *implication set* $S_*(A)$ of A is defined as the infinite union $S_*(A) := \bigcup_{i\geq 0} S_i(A)$. Analogously, $S_*(r) := \bigcup_{i\geq 0} S_i(r)$ and $S_*(a) := \bigcup_{i\geq 0} S_i(a)$.

Since the above definition extends Definition 6 without adding new rules for concept-implication sets $S_*(A)$, Lemma 7 still holds. The following lemma shows that the idea underlying individual-implication sets $S_*(a)$ is also correct in the sense that $A \in S_*(a)$ iff $\mathcal{A} \models_{\mathcal{T}} A(a)$. W.l.o.g. we assume that every individual name $a \in N_{\text{ind}}^{\mathcal{A}}$ has at most one concept assertion $A(a) \in \mathcal{A}$. For every a with $\{A_1(a), A_2(a)\} \subseteq \mathcal{A}$ this can be satisfied by (i) introducing new TBox definitions of the form

$$A_a \sqsubseteq A_1 \sqcap A_2$$
$$A_1 \sqcap A_2 \sqsubseteq A_a,$$

where A_a is a new concept name, and, (ii) modifying $\mathcal{A}$ to $(\mathcal{A}\backslash\{A_1(a), A_2(a)\}) \cup \{A_a(a)\}$. Iterating this modification yields a normalized TBox $\mathcal{T}'$ of linear size in $\mathcal{T}$ with the required property.

Lemma 11 *Let $\mathcal{T}$ be a normalized $\mathcal{ELH}$-TBox over N_{con} and N_{role} and $\mathcal{A}$ an ABox over N_{ind}, $N_{\text{con}}^{\mathcal{T}}$ and $N_{\text{role}}^{\mathcal{T}}$. For every $A_0 \in N_{\text{con}}^{\mathcal{T}}$ and every $a_0 \in N_{\text{ind}}^{\mathcal{A}}$, $A_0 \in S_*(a_0)$ iff $\mathcal{A} \models_{\mathcal{T}} A_0(a_0)$.*

Proof. ($\Rightarrow$) Consider an arbitrary model $\mathcal{I}$ of $\mathcal{T}$ together with $\mathcal{A}$ and arbitrary $A \in N_{\text{con}}^{\mathcal{T}}$ and $a \in \mathcal{A}$ with $A \in S_*(a)$. We prove that $a^{\mathcal{I}} \in A^{\mathcal{I}}$. If $A \in S_*(a)$ then there exist a minimal $n \in \mathbb{N}$ with $A \in S_n(a)$. Proof prove by induction over n.

- ($n = 0$) Then $A = \top$ or $A(a) \in \mathcal{A}$. The claim holds trivially.
- ($n > 0$) There are four cases to distinguish for the rule which caused the inclusion of A into $S_n(a)$.

  (IS1) By induction hypothesis (IH), $a^{\mathcal{I}} \in A_1^{\mathcal{I}}$ for some $A_1 \in N_{\text{con}}^{\mathcal{T},\top}$ with $A_1 \sqsubseteq A \in \mathcal{T}$. Obviously, $a^{\mathcal{I}} \in A^{\mathcal{I}}$ because $\mathcal{I}$ is a model of $\mathcal{T}$.

(**IS2**) Analogous. By IH, $a^{\mathcal{I}} \in A_1^{\mathcal{I}} \cap A_2^{\mathcal{I}}$ for some $A_1, A_2 \in N_{\text{con}}^{\mathcal{T},\top}$ with $A_1 \sqcap A_2 \sqsubseteq A \in \mathcal{T}$. Hence, $a^{\mathcal{I}} \in A^{\mathcal{I}}$ because $\mathcal{I}$ is a model of $\mathcal{T}$.

(**IS3**) By IH, $a \in A_1^{\mathcal{I}}$ for some A_1 and $A_1 \sqsubseteq \exists r.B \in \mathcal{T}$ for some $A_1, B \in N_{\text{con}}^{\mathcal{T},\top}$ and $r \in N_{\text{role}}^{\mathcal{T}}$. Hence, $a^{\mathcal{I}} \in (\exists r.B)^{\mathcal{I}}$. Moreover, there exist $B_1 \in\in N_{\text{con}}^{\mathcal{T},\top}$ and $s \in S_*(r)$ with $B_1 \in S_*(B)$ and $\exists s.B_1 \sqsubseteq A \in \mathcal{T}$. By Lemma 7, $r \sqsubseteq_{\mathcal{T}} s$ and $B \sqsubseteq B_1$, implying $a^{\mathcal{I}} \in (\exists s.B_1)^{\mathcal{I}}$, yielding $a^{\mathcal{I}} \in A^{\mathcal{I}}$.

(**IS4**) Then, $r(a,b) \in \mathcal{A}$, $B \in S_i(b)$, $s \in S_*(r)$, and $\exists r.B \sqsubseteq A \in \mathcal{T}$. By IH $b^{\mathcal{I}} \in B^{\mathcal{I}}$. Hence, $a^{\mathcal{I}} \in (\exists r.B)^{\mathcal{I}}$, implying $a^{\mathcal{I}} \in (\exists s.B)^{\mathcal{I}}$, yielding $a \in A^{\mathcal{I}}$ since $\mathcal{I}$ is a model of $\mathcal{T}$.

($\Leftarrow$) Assume that $A_0 \notin S_*(a_0)$. We construct a model $\mathcal{I}$ of $\mathcal{T}$ and $\mathcal{A}$ where $a^{\mathcal{I}} \notin A^{\mathcal{I}}$. To this end, construct a (possibly infinite) *canonical model* $\mathcal{I}(\mathcal{A})$ of $\mathcal{T}$ together with $\mathcal{A}$ by means of the following definition. $\mathcal{I}(\mathcal{A})$ is defined iteratively starting by $\mathcal{I}_0(\mathcal{A})$. Define $\Delta^{\mathcal{I}_0(\mathcal{A})} := \{x_a \mid a \in N_{\text{ind}}^{\mathcal{A}}\}$, $B^{\mathcal{I}_0(\mathcal{A})} := \{x_a \mid B(a) \in \mathcal{A}\}$, $r^{\mathcal{I}_0(\mathcal{A})} := \{(x_a, x_b) \mid r(a,b) \in \mathcal{A}\}$, and $a^{\mathcal{I}_0(\mathcal{A})} := \{x_a\}$ for all $B \in N_{\text{con}}^{\mathcal{T},\top}$, $r \in N_{\text{role}}^{\mathcal{T}}$, and $a \in N_{\text{ind}}^{\mathcal{A}}$. Note that, w.l.o.g., every $a \in N_{\text{ind}}^{\mathcal{A}}$ is assumed to have at most one assertion $A(a) \in \mathcal{A}$. For $i \geq 0$, the model $\mathcal{I}_{i+1}(\mathcal{A})$ is defined as an extension of $\mathcal{I}_i(\mathcal{A})$ obtained by exhaustive application of the following generation rules.

CM1 If $A \sqsubseteq B \in \mathcal{T}$ then, for every individual $x \in \Delta^{\mathcal{I}_i(\mathcal{A})}$ with $x \in A^{\mathcal{I}_i(\mathcal{A})}$ and $x \notin B^{\mathcal{I}_{i+1}(\mathcal{A})}$, add x to $B^{\mathcal{I}_{i+1}(\mathcal{A})}$;

CM2 If $A \sqcap B \sqsubseteq C \in \mathcal{T}$ then, for every individual $x \in \Delta^{\mathcal{I}_i(\mathcal{A})}$ with $x \in A^{\mathcal{I}_i(\mathcal{A})} \cap B^{\mathcal{I}_i(\mathcal{A})}$ and $x \notin C^{\mathcal{I}_{i+1}(\mathcal{A})}$, add x to $C^{\mathcal{I}_{i+1}(\mathcal{A})}$;

CM3 If $A \sqsubseteq \exists r.B \in \mathcal{T}$ then, for every individual $x \in \Delta^{\mathcal{I}_i(\mathcal{A})}$ with $x \in A^{\mathcal{I}_i(\mathcal{A})}$ for which no r-successor in the interpretation of $B4exists$, introduce a new individual y to $\Delta^{\mathcal{I}_{i+1}(\mathcal{A})}$ and include y into $B^{\mathcal{I}_{i+1}(\mathcal{A})}$ and include (x,y) into $r^{\mathcal{I}_{i+1}(\mathcal{A})}$;

CM4 If $\exists r.A \sqsubseteq B \in \mathcal{T}$ then, for every pair $(x,y) \in s^{\mathcal{I}_i(\mathcal{A})}$ with $s \sqsubseteq_{\mathcal{T}} r$ and $y \in A^{\mathcal{I}_i(\mathcal{A})}$ and $x \notin B^{\mathcal{I}_{i+1}(\mathcal{A})}$, include x into $B^{\mathcal{I}_{i+1}(\mathcal{A})}$;

The above rules are applied fairly, i.e., every rule applicable to already existing elements $x \in \Delta^{\mathcal{I}_i(\mathcal{A})}$ will be applied before applying rules to new elements. The canonical model $\mathcal{I}(\mathcal{A})$ is defined as the infinite union $\mathcal{I}(\mathcal{A}) := \bigcup_{i \geq 0} \mathcal{I}_i(\mathcal{A})$.

Note that the above generation rules are identical to those used in Lemma 7, where a canonical model $\mathcal{I}(A)$ of some $A \in N_{\text{con}}^{\mathcal{T},\top}$ is constructed starting from one individual x_A with $x_A \in A^{\mathcal{I}_0(A)}$. In case of $\mathcal{I}(\mathcal{A})$, we start with one individual x_a for every $a \in N_{\text{ind}}^{\mathcal{A}}$ with $x_a \in A^{\mathcal{I}_0(\mathcal{A})}$ iff $A(a) \in \mathcal{A}$. As every individual name has at most one concept assertion $A(a) \in \mathcal{A}$, the correctness of the construction in Lemma 7 immediately implies $x_a \in A^{\mathcal{I}}$ for every a with $A(a) \in \mathcal{A}$. By definition of $\mathcal{I}_0(\mathcal{A})$, all role assertions $r(a,b) \in \mathcal{A}$ are also satisfied. Hence, $\mathcal{I}(\mathcal{A})$ is a model of $\mathcal{T}$ together with $\mathcal{A}$.

It remains to show that $a_0^{\mathcal{I}(\mathcal{A})} \notin A_0^{\mathcal{I}(\mathcal{A})}$. For every $A \in N_{\text{con}}^{\mathcal{T},\top}$ and $a \in N_{\text{ind}}^{\mathcal{A}}$, $x_a \in A^{\mathcal{I}(\mathcal{A})}$ implies $x_a \in A^{\mathcal{I}_n(\mathcal{A})} \setminus A^{\mathcal{I}_{n-1}(\mathcal{A})}$ for some $n \in \mathbb{N}$. We prove by induction over n that this implies $a \in S_*(A)$.

- $(n = 0)$ $x_a \in A^{\mathcal{I}_0(\mathcal{A})}$ by definition implies $A(a) \in \mathcal{A}$. Hence, $A \in S_0(a)$ by definition of $S_0(a)$.

- $(n > 0)$ Then x_a is added to $\mathcal{I}_n(\mathcal{A})$ by one of the generation Rules **CM1**, **CM2**, or **CM4**.

  (**CM1**) Then there exists a concept name $A_1 \in N_{\text{con}}^{\mathcal{T},\top}$ with $x_a \in A_1^{\mathcal{I}_{n-1}(\mathcal{A})}$ and a GCI $G := A_1 \sqsubseteq A \in \mathcal{T}$. By induction hypothesis (IH) $A_1 \in S_*(a)$, implying $A_1 \in S_m(a) \setminus S_{m-1}(a)$ for some index m. Rule **IS1** with G implies $A \in S_{m+1}(a)$.

  (**CM2**) Analogous. There exist concept names $A_1, A_2 \in N_{\text{con}}^{\mathcal{T},\top}$ and a GCI $G := A_1 \sqcap A_2 \sqsubseteq B \in \mathcal{T}$ with $x_a \in A_1^{\mathcal{I}_{n-1}(\mathcal{A})} \cap A_2^{\mathcal{I}_{n-1}(\mathcal{A})}$. By IH, $\{A_1, A_2\} \subseteq S_*(a)$, implying $\{A_1, A_2\} \subseteq S_m(a) \setminus S_{m-1}(a)$ for some index m. Rule **IS2** with G implies $A \in S_{m+1}(a)$.

  (**CM4**) Then there exists an individual $y \in \Delta^{\mathcal{I}_{n-1}(\mathcal{A})}$ and roles $r, s \in N_{\text{role}}^{\mathcal{T}}$ with $s \sqsubseteq_{\mathcal{T}} r$ such that $(x_a, y) \in s^{\mathcal{I}_{n-1}(\mathcal{A})}$ and $y \in B^{\mathcal{I}_{n-1}(\mathcal{A})}$, where $B \in N_{\text{con}}^{\mathcal{T},\top}$ and $G := \exists r.B \sqsubseteq A \in \mathcal{T}$. By Lemma 7, $r \in S_*(s)$, implying $r \in S_m(s) \setminus S_{m-1}(s)$ for some index m.

  If $y = x_b$ for some $b \in N_{\text{ind}}^{\mathcal{A}}$ then, by definition of the canonical model, $s(a,b) \in \mathcal{A}$, and, by IH, $B \in S_*(b)$, implying $B \in S_{m'}(b) \setminus S_{m'-1}(b)$ for some index m'. Rule **IS4** with G implies $A \in S_{m''+1}(a)$, where m'' denotes the maximum of m and m'.

  If y is not represented by some $b \in N_{\text{ind}}^{\mathcal{A}}$ then it has been added to the model by Rule **CM3**. Hence, there exist concept names $C, D \in N_{\text{con}}^{\mathcal{T},\top}$ and an index $t < n-1$ such that $x_a \in C^{\mathcal{I}_t(\mathcal{A})}$ and $H := C \sqsubseteq \exists s.D \in \mathcal{T}$. By IH, $C \in S_*(a)$. By Rule **CM3**, $(x_a, y) \in s^{\mathcal{I}_{t+1}(\mathcal{A})}$ and $y \in D^{\mathcal{I}_{t+1}(\mathcal{A})}$.

By Lemma 7, $B \in S_*(D)$. Hence, $C \in S_m(a)$ and $B \in S_m(D)$ for some index m. For the least m, Rule IS3 with G and H implies $A \in S_{m+1}(a)$. ∎

Note that the above lemma can be shown without assuming that every individual name has at most one concept assertion in $\mathcal{A}$. Nevertheless, this assumption allows us to exploit the analogy to Lemma 7.

We have shown that the instance problem in $\mathcal{ELH}$ w.r.t. general TBoxes is decidable. The proof of decidability in polynomial time is analogous to Lemma 8: regarding computational complexity, the individual-implication sets $S_*(a)$ have the same properties as concept-implication sets. The new Rule IS4 also does not increase the complexity of computing the sets $S_*(a)$ significantly.

Theorem 12 *The instance problem in $\mathcal{ELH}$ w.r.t. GCIs can be decided in polynomial time.*

5 Conclusion

We have seen how subsumption and instance problem in $\mathcal{ELH}$ w.r.t. general TBoxes can be decided in polynomial time. Moreover, the implication sets computed for one TBox $\mathcal{T}$ can be used to decide *all* subsumptions between defined (or primitive) concepts in $\mathcal{T}$. Hence, classifying $\mathcal{T}$ requires only a single computation of the implication sets for $\mathcal{T}$. The same holds for the instance problem, where a single computation of the relevant implication sets suffices to classify $\mathcal{T}$ and decide *all* instance problems w.r.t. defined (or primitive) concepts occurring in $\mathcal{T}$.

Since subsumption and instance problem remain tractable under the transition from cyclic to general $\mathcal{EL}$-TBoxes, the second natural question is how far the DL can be extended further preserving tractability. Obviously, adding value restrictions makes subsumption NP hard even for the empty TBox [9]. Moreover, it can be shown that adding one of the constructors number restriction, disjunction, or allsome [7] makes subsumption co-NP hard even without GCIs.

It is open, however, whether subsumption and instance problem w.r.t. general TBoxes remain tractable when extending $\mathcal{ELH}$ by inverse roles. Extending our subsumption algorithm by more expressive role constructors

might lead the way to a more efficient reasoning algorithm for the representation language underlying the GALEN [19] terminology, where inverse roles and complex role inclusion axioms can be expressed. While the polynomial upper bound would undoubtedly be exceeded, still a complexity better than EXPTIME might be feasible.

Acknowledgements

My thanks to Carsten Lutz for a multitude of useful remarks and ideas that have greatly influenced this work.

References

[1] F. Baader, J. Hladik, C. Lutz, and F. Wolter. From tableaux to automata for description logics. In Moshe Vardi and Andrei Voronkov, editors, *Proceedings of the 10th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR 2003)*, volume 2850 of *Lecture Notes in Computer Science*, pages 1–32. Springer, 2003.

[2] F. Baader and U. Sattler. An overview of tableau algorithms for description logics. *Studia Logica*, 69:5–40, 2001.

[3] Franz Baader. The instance problem and the most specific concept in the description logic $\mathcal{EL}$ w.r.t. terminological cycles with descriptive semantics. In *Proceedings of the 26th Annual German Conference on Artificial Intelligence, KI 2003*, volume 2821 of *Lecture Notes in Artificial Intelligence*, pages 64–78, Hamburg, Germany, 2003. Springer-Verlag.

[4] Franz Baader. Terminological cycles in a description logic with existential restrictions. In Georg Gottlob and Toby Walsh, editors, *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 325–330. Morgan Kaufmann, 2003.

[5] M. Buchheit, F. M. Donini, and A. Schaerf. Decidable reasoning in terminological knowledge representation systems. *Journal of Artificial Intelligence Research*, 1:109–138, 1993.

[6] R. Cote, D. Rothwell, J. Palotay, R. Beckett, and L. Brochu. The systematized nomenclature of human and veterinary medicine. Technical report, SNOMED International, Northfield, IL, 1993.

[7] Robert Dionne, Eric Mays, and Frank J. Oles. The equivalence of model-theoretic and structural subsumption in description logics. In Ruzena Bajcsy, editor, *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 710–716, San Mateo, California, 1993. Morgan Kaufmann.

[8] F.M. Donini. Complexity of reasoning. In Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, pages 96–136. Cambridge University Press, 2003.

[9] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi, Bernhard Hollunder, Werner Nutt, and Alberto Spaccamela. The complexity of existential quantification in concept languages. *Artificial Intelligence*, 53(2–3):309–327, 1992.

[10] Robert Givan, David A. McAllester, Carl Witty, and Dexter Kozen. Tarskian set constraints. *Information and Computation*, 174(2):105–131, 2002.

[11] Volker Haarslev and Ralf Möller. RACER system description. *Lecture Notes in Computer Science*, 2083:701–712, 2001.

[12] Ian Horrocks, Alan L. Rector, and Carole A. Goble. A description logic based schema for the classification of medical data. In *Knowledge Representation Meets Databases*, 1996.

[13] Ian Horrocks, Ulrike Sattler, and Stephan Tobies. Practical reasoning for expressive description logics. In Harald Ganzinger, David McAllester, and Andrei Voronkov, editors, *Proceedings of the 6th International Conference on Logic for Programming and Automated Reasoning (LPAR'99)*, number 1705 in Lecture Notes in Artificial Intelligence, pages 161–180. Springer-Verlag, September 1999.

[14] Ian R. Horrocks. Using an expressive description logic: FaCT or fiction? In Anthony G. Cohn, Lenhart Schubert, and Stuart C. Shapiro, editors,

KR'98: Principles of Knowledge Representation and Reasoning, pages 636–645. Morgan Kaufmann, San Francisco, California, 1998.

[15] Yevgeny Kazakov and Hans De Nivelle. Subsumption of concepts in $\mathcal{FL}_0$ for (cyclic) terminologies with respect to descriptive semantics is pspace-complete. In *Proceedings of the 2003 International Workshop on Description Logics (DL2003)*, CEUR-WS, 2003.

[16] C. Lutz. Complexity of terminological reasoning revisited. In *Proceedings of the 6th International Conference on Logic for Programming and Automated Reasoning LPAR'99*, Lecture Notes in Artificial Intelligence, pages 181–200. Springer-Verlag, September 6 – 10, 1999.

[17] B. Nebel. Terminological cycles: Semantics and computational properties. In J. F. Sowa, editor, *Principles of Semantic Networks: Explorations in the Representation of Knowledge*, pages 331–361. Morgan Kaufmann Publishers, San Mateo (CA), USA, 1991.

[18] A. Rector. Medical informatics. In Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, pages 406–426. Cambridge University Press, 2003.

[19] A. Rector, S. Bechhofer, C. A. Goble, I. Horrocks, W. A. Nowlan, and W. D. Solomon. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, 9:139–171, 1997.

[20] A. Rector, W. Nowlan, and A. Glowinski. Goals for concept representation in the GALEN project. In *Proceedings of the 17th annual Symposium on Computer Applications in Medical Care, Washington, USA, SCAMC*, pages 414–418, 1993.

[21] K. Spackman. Normal forms for description logic expressions of clinical concepts in SNOMED RT. *Journal of the American Medical Informatics Association*, (Symposium Supplement), 2001.