The Logical Difference Problem for Description Logic Terminologies

Boris Konev, Dirk Walther, and Frank Wolter

University of Liverpool, Liverpool, UK
{konev, dwalther, wolter}@liverpool.ac.uk

Abstract. We consider the problem of computing the logical difference between distinct versions of description logic terminologies. For the lightweight description logic $\mathcal{EL}$, we present a tractable algorithm which, given two terminologies and a signature, outputs a set of concepts, which can be regarded as the logical difference between the two terminologies. As a consequence, the algorithm can also decide whether they imply the same concept implications in the signature. A prototype implementation CEX of this algorithm is presented and experimental results based on distinct versions of SNOMED CT, the Systematized Nomenclature of Medicine, Clinical Terms, are discussed. Finally, results regarding the relation to uniform interpolants and possible extensions to more expressive description logics are presented.

1 Introduction

The standard diff operation for text files is an indispensable tool for comparing different versions of texts, and similar operations are available to software engineers comparing distinct versions of code produced in collaborative software projects. As observed, e.g., in [14], such a purely syntactic diff operation is hardly useful if the text consists of a set of axioms of an ontology. In this case, one is usually not interested in a comparison of the syntactic form of axioms, but in the consequences that the ontologies have. The authors of [14] present a number of heuristic rules to address this problem and develop a diff operator for ontologies. Except theoretical results in [12, 13, 9], we are not aware of any logic-based approach to computing the logical diff of ontologies.

Our formalisation of the logical difference problem is based on the observation that when comparing distinct versions of ontologies one should take into account their signatures. In fact, the interesting differences between ontologies are those formulated in their shared signature (or even subsets thereof), and not those involving symbols used only in one of the two ontologies. Thus, the proposed notion of logical difference is based on the notion of Σ-*entailment*: an ontology $\mathcal{T}$ Σ-entails an ontology $\mathcal{T}'$ for a signature Σ, if for all concept implications $C \sqsubseteq D$ in Σ, $\mathcal{T}' \models C \sqsubseteq D$ implies $\mathcal{T} \models C \sqsubseteq D$. If $\mathcal{T}$ and $\mathcal{T}'$ mutually Σ-entail each other, then they are called Σ-*inseparable*. By taking Σ as the set of shared symbols of $\mathcal{T}$ and $\mathcal{T}'$, Σ-inseparability means that $\mathcal{T}$ and $\mathcal{T}'$ are not

distinguishable by means of concept implications in their shared signature. In this case, their logical difference will be regarded as empty.

We show that deciding Σ-entailment is tractable for $\mathcal{EL}$-terminologies, i.e., sets of possibly cyclic concept definitions in the lightweight description logic $\mathcal{EL}$; see [1, 10]. Observe that for ontologies formulated as general TBoxes in description logics, the computational complexity of deciding Σ-entailment is by at least one exponential harder than the deduction problem, e.g., it is 2ExpTime-complete for expressive description logics such as $\mathcal{ALC}$, $\mathcal{ALCQ}$, and $\mathcal{ALCQI}$ [6, 12] and ExpTime-complete for $\mathcal{EL}$ itself [13]. Moreover, even in such simple formalisms as acyclic propositional Horn Logic Σ-entailment is co-NP-complete [5].

In applications, it is not enough to decide whether two ontologies are logically different, but an informative list of differences is required. We show that for any concept implication $C \sqsubseteq D$ in the logical difference between two $\mathcal{EL}$-terminologies, there exist subconcepts C' and D' of C and D, respectively, such that $C' \sqsubseteq D'$ is in the logical difference and C' or D' is a concept name. Thus, listing the set of all concept names involved in such implications appears to be an informative approximation of the logical difference between two $\mathcal{EL}$-terminologies. This list is empty if, and only if, there is no logical difference between the two terminologies.

The system CEX implements, by employing a dynamic programming approach, the algorithm deciding Σ-entailment and lists the set of logical differences described above for acyclic $\mathcal{EL}$-terminologies. We present a variety of experiments in which CEX is applied to different versions of Snomed ct, the Systematized Nomenclature of Medicine, Clinical Terms. This terminology comprises ∼0.4 million terms and underlies the systematised medical terminology used in the health systems of the US, the UK, and other countries [17].

Finally, we discuss an alternative approach to deciding Σ-entailment using uniform interpolants and explore the complexity of corresponding reasoning problems for acyclic $\mathcal{ALC}$-terminologies.

Detailed proofs are provided in the technical report [11].

2 Preliminaries

Let $\mathsf{N_C}$ and $\mathsf{N_R}$ be countably infinite and disjoint sets of *concept names* and *role names*, respectively. In the description logic $\mathcal{EL}$, *concepts* C are built according to the syntax rule

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

where A ranges over $\mathsf{N_C}$, r ranges over $\mathsf{N_R}$, and C, D range over concepts. The semantics of concepts is defined by means of *interpretations* $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$, where the interpretation *domain* $\Delta^\mathcal{I}$ is a non-empty set, and $\cdot^\mathcal{I}$ is a function mapping each concept name A to a subset $A^\mathcal{I}$ of $\Delta^\mathcal{I}$ and each role name $r^\mathcal{I}$ to a binary relation $r^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I}$. The function $\cdot^\mathcal{I}$ is inductively extended to arbitrary concepts by setting $\top^\mathcal{I} := \Delta^\mathcal{I}$, $(C \sqcap D)^\mathcal{I} := C^\mathcal{I} \cap D^\mathcal{I}$, and $(\exists r.C)^\mathcal{I} := \{d \in \Delta^\mathcal{I} \mid \exists e \in C^\mathcal{I} : (d, e) \in r^\mathcal{I}\}$.

A *general TBox* is a finite set of *axioms*, where an axiom can be either a *concept inclusion (CI)* $C \sqsubseteq D$ or a *concept equality (CE)* $C \equiv D$, where C, D are concepts. An interpretation $\mathcal{I}$ *satisfies* a CI $C \sqsubseteq D$ (written $\mathcal{I} \models C \sqsubseteq D$) if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$; it *satisfies* a CE $C \equiv D$ (written $\mathcal{I} \models C \equiv D$) if $C^{\mathcal{I}} = D^{\mathcal{I}}$. $\mathcal{I}$ is a *model* of a general TBox $\mathcal{T}$ if it satisfies all axioms in $\mathcal{T}$. We write $\mathcal{T} \models C \sqsubseteq D$ ($\mathcal{T} \models C \equiv D$) if every model of $\mathcal{T}$ satisfies $C \sqsubseteq D$ ($C \equiv D$, respectively).

Our main concern in this paper are *terminologies*, i.e., general TBoxes $\mathcal{T}$ satisfying the following two conditions:

- $\mathcal{T}$ consists of CEs of the form $A \equiv C$ (*concept definitions*) and CIs of the form $A \sqsubseteq C$ (*primitive concept definitions*) only, where A is a concept name;
- no concept name occurs more than once on the left hand side of an axiom in $\mathcal{T}$.

Define the relation $\prec_{\mathcal{T}}$ between concept names by setting $A \prec_{\mathcal{T}} B$ if there exists an axiom of the form $A \equiv C$ or $A \sqsubseteq C$ in $\mathcal{T}$ such that B occurs in C. A terminology $\mathcal{T}$ is called *acyclic* if the transitive closure $\prec_{\mathcal{T}}^*$ of $\prec_{\mathcal{T}}$ is irreflexive.

A *signature* Σ is a finite subset of $\mathsf{N_C} \cup \mathsf{N_R}$. The signature $\mathsf{sig}(C)$ ($\mathsf{sig}(\alpha)$, $\mathsf{sig}(\mathcal{T})$) of a concept C (axiom α, terminology $\mathcal{T}$) is the set of concept and role names which occur in C (α, $\mathcal{T}$, respectively). If $\mathsf{sig}(C) \subseteq \Sigma$, we also call C a *Σ-concept* and similarly for axioms and terminologies.

Definition 1 (Σ-difference, Σ-entailment). *Let $\mathcal{T}$ and $\mathcal{T}'$ be terminologies and Σ a signature. The Σ-difference, $\mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$, between $\mathcal{T}$ and $\mathcal{T}'$ is defined as*

$$\mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}') = \{C \sqsubseteq D \mid \mathcal{T} \not\models C \sqsubseteq D \text{ and } \mathcal{T}' \models C \sqsubseteq D \text{ and } \mathsf{sig}(C \sqsubseteq D) \subseteq \Sigma\}.$$

$\mathcal{T}$ Σ-entails $\mathcal{T}'$ if, and only if, $\mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}') = \emptyset$. $\mathcal{T}$ and $\mathcal{T}'$ are called Σ-inseparable if $\mathcal{T}$ and $\mathcal{T}'$ Σ-entail each other.

Example 1. Observe that, in some cases, $\mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$ only contains concept implications of at least exponential size, even for acyclic terminologies. To start with, let $\mathcal{T} = \emptyset$,

$$\mathcal{T}' = \{A_0 \sqsubseteq B_0, A_1 \equiv B_n\} \cup \{B_{i+1} \equiv \exists r.B_i \sqcap \exists s.B_i \mid 0 \leq i < n\},$$

and $\Sigma = \{A_0, A_1, r, s\}$. Then $\mathcal{T}'$ is not Σ-entailed by $\mathcal{T}$, and a minimal implication of the form $C \sqsubseteq A_1$ in $\mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$ is given by $C_n \sqsubseteq A_1$, where $C_0 = A_0$ and $C_{i+1} = \exists r.C_i \sqcap \exists s.C_i$, for $i \geq 0$. Clearly, C_n is of exponential size. Observe, however, that there exist much smaller implications than $C_n \sqsubseteq A_1$ in $\mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$. Namely, $A_1 \sqsubseteq \exists r.\top$, $A_1 \sqsubseteq \exists s.\top$, $A_1 \sqsubseteq \exists r.\top \sqcap \exists s.\top$, etc. To avoid this type of implications in $\mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$ replace $\mathcal{T}$ by

$$\mathcal{T}_0 = \{A_1 \sqsubseteq F_0\} \cup \{F_i \sqsubseteq \exists r.F_{i+1} \sqcap \exists s.F_{i+1} \mid 0 \leq i < n\}.$$

Then one can easily see that $C_n \sqsubseteq A_1$ is the smallest implication in $\mathsf{Diff}_\Sigma(\mathcal{T}_0, \mathcal{T}')$. Observe, however, that if we use structure sharing and define the size of C_n as the number of its subconcepts, then C_n is only of polynomial size.

$$\frac{}{C \sqsubseteq C}\ (\textsc{Ax}) \quad \frac{}{C \sqsubseteq \top}\ (\textsc{AxTop}) \quad \frac{C \sqsubseteq E}{C \sqcap D \sqsubseteq E}\ (\textsc{AndL1}) \quad \frac{D \sqsubseteq E}{C \sqcap D \sqsubseteq E}\ (\textsc{AndL2})$$

$$\frac{C \sqsubseteq E \quad C \sqsubseteq D}{C \sqsubseteq D \sqcap E}\ (\textsc{AndR}) \quad \frac{C \sqsubseteq D}{\exists r.C \sqsubseteq \exists r.D}\ (\textsc{Ex})$$

$$\frac{C_A \sqsubseteq D}{A \sqsubseteq D}\ (\textsc{DefL}) \quad \frac{D \sqsubseteq C_A}{D \sqsubseteq A}\ (\textsc{DefR}) \quad \text{where } A \equiv C_A \in \mathcal{T}$$

$$\frac{C_A \sqsubseteq D}{A \sqsubseteq D}\ (\textsc{PDefL}) \quad \text{where } A \sqsubseteq C_A \in \mathcal{T}$$

Fig. 1. Gentzen-style proof system for $\mathcal{EL}$ terminologies.

Observe that if $\mathcal{T}$ Σ-entails $\mathcal{T}'$, then $\mathcal{T}$ Σ'-entails $\mathcal{T}'$ for any Σ' with $\Sigma' \cap \mathsf{sig}(\mathcal{T}') \subseteq \Sigma$. This follows immediately from the following interpolation result [16].

Theorem 1. *$\mathcal{EL}$ has the interpolation property, i.e., if $\mathcal{T} \models C \sqsubseteq D$, then there exists a finite set $\mathcal{T}_0$ of CIs with $\mathsf{sig}(\mathcal{T}_0) \subseteq \mathsf{sig}(\mathcal{T}) \cap \mathsf{sig}(C \sqsubseteq D)$ such that $\mathcal{T} \models \mathcal{T}_0$ and $\mathcal{T}_0 \models C \sqsubseteq D$.*

3 Basic properties of $\mathcal{EL}$

We derive basic properties of $\mathcal{EL}$ from the Gentzen-style sequent calculus of Hofmann [10], see Figure 1.[1] The basic calculus of [10] considers $\mathcal{EL}$ without the constant $\top$ and for terminologies without primitive concept definitions. To take care of $\top$, we have added the rule (AxTop), and (PDefL) is the rule representing axioms of the form $A \sqsubseteq C$. Cut-elimination, completeness, and correctness can now be proved by a straightforward extension of the proof in [10]. For a terminology $\mathcal{T}$ and concepts C, D, we write $\mathcal{T} \vdash C \sqsubseteq D$ iff there exists a proof of $C \sqsubseteq D$ in the calculus of Figure 1.

Theorem 2 (Hofmann). *For all terminologies $\mathcal{T}$ and concepts C, D, it holds that $\mathcal{T} \models C \sqsubseteq D$ if, and only if, $\mathcal{T} \vdash C \sqsubseteq D$.*

We apply this calculus to derive a description of the syntactic form of concepts C such that $\mathcal{T} \models C \sqsubseteq D$, where D is not equivalent to a conjunction. Call a concept name A *primitive in* $\mathcal{T}$ if A does not occur on the left hand side of an axiom in $\mathcal{T}$. Call A *pseudo-primitive in* $\mathcal{T}$ if it is primitive in $\mathcal{T}$ or occurs on the left hand side of primitive concept definitions in $\mathcal{T}$. In what follows, we say that a concept F is a conjunction of concepts if $F = \bigsqcap_{D \in X} D$, for a set X of concepts. Any $D \in X$ is then called a *conjunct* of F and, if D is a concept name,

[1] Alternatively, one could start from the model-theoretic analysis of $\mathcal{EL}$ terminologies in [1].

then it is called an *atomic conjunct* of F. We sometimes write $D \in F$ instead of $D \in X$ and if X is empty, then F denotes the concept $\top$.

Lemma 1. *Let $\mathcal{T}$ be a terminology and $C = F \sqcap \bigsqcap_{(r,D)\in Q} \exists r.D$, where F is a conjunction of concept names and Q is a set of pairs (r, D) in which r is a role and D a concept.*

1. *If $\mathcal{T} \models C \sqsubseteq A$ for an A which is pseudo-primitive in $\mathcal{T}$, then $\mathcal{T} \models B \sqsubseteq A$, for some atomic conjunct B of F.*
2. *If $\mathcal{T} \models C \sqsubseteq \exists s.C_0$, then*
   - *$\mathcal{T} \models B \sqsubseteq \exists s.C_0$, for some atomic conjunct B of F, or*
   - *there exists $(r, D) \in Q$ such that $r = s$ and $\mathcal{T} \models D \sqsubseteq C_0$.*

Proof. We use Theorem 2 and prove Point 1. Point 2 is proved similarly. Let $\mathcal{T} \vdash C \sqsubseteq A$, where A is pseudo-primitive in $\mathcal{T}$. Let $\mathcal{D}$ be a proof of $C \sqsubseteq A$. Note that, since A is pseudo-primitive in $\mathcal{T}$, $\mathcal{D}$ can only end with one of Ax, AndL1, AndL2, DefL, or PDefL. We show that then $\mathcal{T} \vdash B \sqsubseteq A$, for some conjunct B of F, by induction on the number n of conjuncts in C.

The base case of $n = 1$ is trivial: $\mathcal{D}$ can only end with one of Ax, PDefL, or DefL; so, C is a concept name itself.

Assume $n > 1$. Then $\mathcal{D}$ can only end with one of AndL1 or AndL2. In any case, there exists a conjunct C' of C such that $\mathcal{T} \vdash C' \sqsubseteq A$ and C' contains less conjuncts than C. By induction, there exists a concept name B which is a conjunct in C' such that $\mathcal{T} \vdash B \sqsubseteq A$. Note now that B is also a conjunct of C.

We apply Lemma 1 to show that if $\mathcal{T}$ does not Σ-entail $\mathcal{T}'$, then there exists $C \sqsubseteq D \in \mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$ such that C or D is a concept name.

Lemma 2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be terminologies and Σ a signature. If $C \sqsubseteq D \in \mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$, then there exist subconcepts C' and D' of C and D, respectively, such that $C' \sqsubseteq D' \in \mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$ and $C' \sqsubseteq D'$ is of the form $A \sqsubseteq \exists r.D_0$ or $C_0 \sqsubseteq A$, where A is a concept name.*

Proof. Let $C \sqsubseteq D \in \mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$. Then $D \neq \top$ because otherwise $\mathcal{T} \models C \sqsubseteq D$. If $D = D_1 \sqcap D_2$, then one of $C \sqsubseteq D_i$, $i = 1, 2$, is in the Σ-difference. If $D = \exists r.D_1$ then, by Lemma 1, either there exists a subconcept A of C, A a concept name, such that $A \sqsubseteq D$ is in the Σ-difference, or there exists a subconcept $\exists r.C_1$ of C, such that $C_1 \sqsubseteq D_1$ is in the Σ-difference. Simplify $C \sqsubseteq D$ until none of these simplification rules is applicable. The resulting CI is as required.

4 Deciding Σ-entailment: theory

By Lemma 2, to decide Σ-entailment, it is sufficient to decide whether the set $\mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$ contains Σ-implications of the form $C \sqsubseteq A$ or $A \sqsubseteq D$, where A is a concept name. The latter problem is decidable in polynomial time already for general $\mathcal{EL}$-TBoxes [13]. So, in what follows we concentrate on Σ-implications of the form $C \sqsubseteq A$. We first transform $\mathcal{T}$ into a *normalised* terminology. A concept

For $A \in \mathsf{N_C}$,

- if A is pseudo-primitive in $\mathcal{T}$, then

$$\mathsf{noimply}_{\mathcal{T},\Sigma}(A) = \{\xi_A\}, \quad \mathsf{Noimply}_{\mathcal{T},\Sigma}(A) = \{\xi_A \sqsubseteq \bigsqcap_{A' \in (\Sigma \setminus \mathsf{pre}^{\Sigma}_{\mathcal{T}}(A))} A' \sqcap \mathsf{All}_\Sigma\};$$

- if A is conjunctive in $\mathcal{T}$ and $A \equiv F \in \mathcal{T}$, then

$$\mathsf{noimply}_{\mathcal{T},\Sigma}(A) = \{\xi_B \mid B \in F\}, \quad \mathsf{Noimply}_{\mathcal{T},\Sigma}(A) = \emptyset$$

- if $A \equiv \exists r.B \in \mathcal{T}$, then $\mathsf{noimply}_{\mathcal{T},\Sigma}(A) = \{\xi_A\}$ and $\mathsf{Noimply}_{\mathcal{T},\Sigma}(A) = \{\alpha_A\}$, where

$$\alpha_A = \xi_A \sqsubseteq (\bigsqcap_{A' \in (\Sigma \setminus \mathsf{pre}^{\Sigma}_{\mathcal{T}}(A))} A') \sqcap (\bigsqcap_{r \neq s \in \Sigma} \exists s.(\bigsqcap_{A' \in \Sigma} A' \sqcap \mathsf{All}_\Sigma)) \sqcap \bigsqcap_{\xi \in \mathsf{noimply}_{\mathcal{T},\Sigma}(B)} \exists r.\xi.$$

Fig. 2. Computing $\mathsf{Noimply}_{\mathcal{T},\Sigma}(A)$ and $\mathsf{noimply}_{\mathcal{T},\Sigma}(A)$

name A is called *non-conjunctive in* $\mathcal{T}$ if it is pseudo-primitive in $\mathcal{T}$ or has a definition of the form $A \equiv \exists r.C \in \mathcal{T}$. Otherwise A is called *conjunctive in* $\mathcal{T}$. A terminology $\mathcal{T}$ is normalised if it consists of axioms of the following form:

- $A \equiv \exists r.B$ or $A \sqsubseteq \exists r.B$, where B is a concept name;
- $A \equiv F$ or $A \sqsubseteq F$, where F is a (possibly empty) conjunction of concept names such that every conjunct B of F is non-conjunctive in $\mathcal{T}$.

Normalised terminologies in the sense defined above are a minor modification of normalised terminologies as defined in [1]. Say that two interpretations $\mathcal{I}$ and $\mathcal{J}$ *coincide on a signature* Σ, in symbols $\mathcal{I}|_\Sigma = \mathcal{J}|_\Sigma$, if $\Delta^{\mathcal{I}} = \Delta^{\mathcal{J}}$ and $X^{\mathcal{I}} = X^{\mathcal{J}}$ for all $X \in \Sigma$.

Lemma 3. *For every terminology $\mathcal{T}$, one can construct in polynomial time a normalised terminology $\mathcal{T}'$ of polynomial size in $|\mathcal{T}|$ such that $\mathsf{sig}(\mathcal{T}) \subseteq \mathsf{sig}(\mathcal{T}')$, $\mathcal{T}' \models \mathcal{T}$, and for every model $\mathcal{I}$ of $\mathcal{T}$ there exists a model $\mathcal{J}$ of $\mathcal{T}'$ which coincides with $\mathcal{I}$ on Σ. Moreover, $\mathcal{T}'$ is acyclic if $\mathcal{T}$ is acyclic.*

The proof is a straightforward modification of the proof in [1]. From now on we will work with normalised terminologies only.

Intuitively, to decide whether there exists $C \sqsubseteq A \in \mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$, we want to construct the most specific[2] Σ-concept C_A such that $\mathcal{T} \not\models C_A \sqsubseteq A$. Then there exists *some* Σ-concept C such that $C \sqsubseteq A \in \mathsf{Diff}_\Sigma(\mathcal{T}, \mathcal{T}')$ if, and only if, $\mathcal{T}' \models C_A \sqsubseteq A$. Unfortunately, most specific Σ-concepts with this property do not always exist and, therefore (and also to enable structure sharing), we use an additional terminology. We use the following sets and axiom:

- $\Sigma^{\mathsf{fresh}} = \{\mathsf{All}_\Sigma\} \cup \{\xi_A \mid A \in \mathsf{N_C} \text{ non-conjunctive in } \mathcal{T}\}$, where All_Σ and each ξ_A are fresh concept names not occurring in $\Sigma \cup \mathsf{sig}(\mathcal{T})$;

[2] Recall that a concept C is more specific than a concept D if $\models C \sqsubseteq D$.

– α denotes the concept inclusion $\mathsf{All}_\Sigma \sqsubseteq \bigsqcap_{r\in\Sigma} \exists r.(\bigsqcap_{A'\in\Sigma} A' \sqcap \mathsf{All}_\Sigma)$;
– $\mathsf{pre}^\Sigma_\mathcal{T}(A) = \{B \in \Sigma \mid \mathcal{T} \models B \sqsubseteq A\}$, for $A \in \mathsf{N_C}$. These sets can be computed in polynomial time [1].

Theorem 3. *Let $\mathcal{T}$ be a normalised terminology and Σ a signature. The terminologies $\mathsf{Noimply}_{\mathcal{T},\Sigma}(A)$ and sets of concepts names $\mathsf{noimply}_{\mathcal{T},\Sigma}(A)$ are constructed, in polynomial time, in Figure 2. Set*

$$\mathsf{Noimply}_{\mathcal{T},\Sigma} = \{\alpha\} \cup \bigcup_{A\in\Sigma\cup\mathsf{sig}(\mathcal{T})} \mathsf{Noimply}_{\mathcal{T},\Sigma}(A).$$

The following conditions are equivalent, for every concept name $A \in \Sigma \cup \mathsf{sig}(\mathcal{T})$ and terminology $\mathcal{T}'$ with $\mathsf{sig}(\mathcal{T}') \cap \Sigma^{\mathsf{fresh}} = \emptyset$:

– *there exists a Σ-concept C with $\mathcal{T}' \models C \sqsubseteq A$ and $\mathcal{T} \not\models C \sqsubseteq A$;*
– *$\mathcal{T}' \cup \mathsf{Noimply}_{\mathcal{T},\Sigma} \models \xi \sqsubseteq A$, for some $\xi \in \mathsf{noimply}_{\mathcal{T},\Sigma}(A)$.*

Observe that, in Theorem 3, $\mathsf{Noimply}_{\mathcal{T},\Sigma}$ and $\mathsf{noimply}_{\mathcal{T},\Sigma}(A)$ do not depend on $\mathcal{T}'$. Thus, once they have been constructed, they can be used to check the existence of concept implications $C \sqsubseteq A \in \mathsf{Diff}_\Sigma(\mathcal{T},\mathcal{T}')$ for arbitrary terminologies $\mathcal{T}'$. It is worth noting as well that the proof of Theorem 3 will show that the result holds for arbitrary general TBoxes $\mathcal{T}'$ formulated in description logics which are fragments of first-order logic, and, indeed, for $\mathcal{T}'$ any first-order theory. In this case, Theorem 3 provides a reduction of checking whether there exists $C \sqsubseteq A \in \mathsf{Diff}_\Sigma(\mathcal{T},\mathcal{T}')$ to deduction in the language of $\mathcal{T}'$.

Example 2. Let $\mathcal{T} = \{A \equiv \exists r.B, B \equiv \exists r.A\}$ and $\Sigma = \{r, A, B\}$. Then we have $\mathsf{noimply}_{\mathcal{T},\Sigma}(A) = \{\xi_A\}$ and $\mathsf{Noimply}_{\mathcal{T},\Sigma} = \{\xi_A \sqsubseteq B \sqcap \exists r.\xi_B,\ \xi_B \sqsubseteq A \sqcap \exists r.\xi_A\}$. Intuitively, $\{\xi_A\} \cup \mathsf{Noimply}_{\mathcal{T},\Sigma}$ stands for the "infinitary" most specific Σ-concept not subsumed by A relative to $\mathcal{T}$.

In the remainder of this section we prove Theorem 3. To this end, we first prove an "infinitary" version of Theorem 3 by associating with every concept name A a sequence $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$, $n \geq 0$, of sets of Σ-concepts such that the following holds:

C1. $\mathcal{T} \not\models C \sqsubseteq A$, for all $n \geq 0$ and for all $C \in \mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$.
C2. For all Σ-concepts D, if $\mathcal{T} \not\models D \sqsubseteq A$, then $\models C \sqsubseteq D$ for some $C \in \mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$, where n is the *role-depth* $\mathsf{depth}(D)$ of D (i.e., the number of nestings of existential restrictions in D).[3]

The sets $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$ are defined in Figure 3. Observe that $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$ is well-defined because in the definition $A \equiv F \in \mathcal{T}$ of a conjunctive concept name A no conjunctive concept name occurs. This observation will also be used in the inductive proofs below.

[3] More precisely $\mathsf{depth}(A) = 0$, $\mathsf{depth}(C_1 \sqcap C_2) = \max\{\mathsf{depth}(C_1), \mathsf{depth}(C_2)\}$, and $\mathsf{depth}(\exists r.D) = \mathsf{depth}(D) + 1$.

Set, inductively, $\mathsf{all}^0_\Sigma = \top$ and $\mathsf{all}^{n+1}_\Sigma = \bigsqcap_{r\in\Sigma} \exists r.(\bigsqcap_{A'\in\Sigma} A' \sqcap \mathsf{all}^n_\Sigma)$. Define $\mathsf{noimply}^0_{\mathcal{T},\Sigma}(A)$ as follows:

- if A is non-conjunctive in $\mathcal{T}$, then $\mathsf{noimply}^0_{\mathcal{T},\Sigma}(A) = \{\bigsqcap_{A'\in\Sigma\setminus\mathsf{pre}^\Sigma_\mathcal{T}(A)} A'\}$;
- if A is conjunctive and $A \equiv F \in \mathcal{T}$, then $\mathsf{noimply}^0_{\mathcal{T},\Sigma}(A) = \bigcup_{B\in F} \mathsf{noimply}^0_{\mathcal{T},\Sigma}(B)$;

and define, inductively, $\mathsf{noimply}^{n+1}_{\mathcal{T},\Sigma}(A)$ by

- if A is pseudo-primitive in $\mathcal{T}$, then $\mathsf{noimply}^{n+1}_{\mathcal{T},\Sigma}(A) = \{\bigsqcap_{A'\in(\Sigma\setminus\mathsf{pre}^\Sigma_\mathcal{T}(A))} A' \sqcap \mathsf{all}^{n+1}_\Sigma\}$.
- If A is conjunctive and $A \equiv F \in \mathcal{T}$, then $\mathsf{noimply}^{n+1}_{\mathcal{T},\Sigma}(A) = \bigcup_{B\in F} \mathsf{noimply}^{n+1}_{\mathcal{T},\Sigma}(B)$.
- If $A \equiv \exists r.B \in \mathcal{T}$, then $\mathsf{noimply}^{n+1}_{\mathcal{T},\Sigma}(A) = \{C^{n+1}_{\Sigma,\mathcal{T}}\}$, where

$$C^{n+1}_{\Sigma,\mathcal{T}} = \Big(\bigsqcap_{A'\in(\Sigma\setminus\mathsf{pre}^\Sigma_\mathcal{T}(A))} A' \sqcap \Big(\bigsqcap_{r\neq s\in\Sigma} \exists s.\Big(\bigsqcap_{A'\in\Sigma} A' \sqcap \mathsf{all}^n_\Sigma\Big)\Big) \sqcap \bigsqcap_{E\in\mathsf{noimply}^n_{\mathcal{T},\Sigma}(B)} \exists r.E.$$

Fig. 3. Computing $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$

Example 3. For the terminology $\mathcal{T}$ and signature Σ from Example 2, we have $\mathsf{noimply}^0_{\mathcal{T},\Sigma}(A) = \{B\}$, $\mathsf{noimply}^1_{\mathcal{T},\Sigma}(A) = \{B \sqcap \exists r.A\}$, $\mathsf{noimply}^2_{\mathcal{T},\Sigma}(A) = \{B \sqcap \exists r.(A \sqcap \exists r.B)\}$, etc. Thus, intuitively, $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$ is the unfolding up to depth n of ξ_A relative to $\mathsf{Noimply}_{\mathcal{T},\Sigma}$.

Lemma 4. *Let $\mathcal{T}$ be a normalised terminology, signature Σ, and $A \in \mathsf{N_C}$. The sets $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$ satisfy conditions **C1** and **C2** above.*

Proof. We start with the proof of **C1**. Assume first that A is pseudo-primitive in $\mathcal{T}$. Then $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$ consists of $C = \bigsqcap_{A'\in(\Sigma\setminus\mathsf{pre}^\Sigma_\mathcal{T}(A))} A' \sqcap \mathsf{all}^n_\Sigma$. By Lemma 1, $\mathcal{T} \not\models C \sqsubseteq A$ because the only atomic conjuncts of C are in $\Sigma \setminus \mathsf{pre}^\Sigma_\mathcal{T}(A)$.

We now prove **C1** for concept names A which are not pseudo-primitive in $\mathcal{T}$. The proof is by induction on n. For $n = 0$ and $A \equiv \exists r.B \in \mathcal{T}$ the claim follows again from Lemma 1 and the observation that $B' \in \mathsf{pre}^\Sigma_\mathcal{T}(A)$ if, and only if, $\mathcal{T} \models B' \sqsubseteq \exists r.B$. For $n = 0$ and A conjunctive with $A \equiv F \in \mathcal{T}$, **C1** follows since it has been proved for all conjuncts of F and $\mathcal{T} \not\models C \sqsubseteq A$ if, and only if, there exists an atomic conjunct B of F such that $\mathcal{T} \not\models C \sqsubseteq B$.

For the induction step, assume **C1** has been proved for $n \geq 0$.

Let $A \equiv \exists r.B \in \mathcal{T}$ and let $C^{n+1}_{\mathcal{T},\Sigma}$ be the only element of $\mathsf{noimply}^{n+1}_{\mathcal{T},\Sigma}(A)$. Assume $\mathcal{T} \models C^{n+1}_{\mathcal{T},\Sigma} \sqsubseteq A$. By Lemma 1 there are two cases:

- $\mathcal{T} \models \bigsqcap_{A'\in(\Sigma\setminus\mathsf{pre}^\Sigma_\mathcal{T}(A))} A' \sqsubseteq \exists r.B$. This is excluded, by Lemma 1.
- There exists $E \in \mathsf{noimply}^n_{\mathcal{T},\Sigma}(B)$ such that $\mathcal{T} \models E \sqsubseteq B$. This is excluded by the IH.

We have derived a contradiction. The case $A \equiv F \in \mathcal{T}$, A conjunctive in $\mathcal{T}$, is considered similarly to the case $n = 0$ and left to the reader.

We come to the proof of **C2**. The proof is by induction on n. Let $n = 0$ and assume first that A is non-conjunctive. Let D be a Σ-concept with $\mathsf{depth}(D) = 0$

and $\mathcal{T} \not\models D \sqsubseteq A$. Then all conjuncts of D are in $\Sigma \setminus \mathsf{pre}^{\Sigma}_{\mathcal{T}}(A)$ and we obtain $\models \bigsqcap_{A' \in \Sigma \setminus \mathsf{pre}^{\Sigma}_{\mathcal{T}}(A)} A' \sqsubseteq D$. Now assume A is conjunctive in $\mathcal{T}$ and $A \equiv F \in \mathcal{T}$. Let D be a Σ-concept with $\mathsf{depth}(D) = 0$ and $\mathcal{T} \not\models D \sqsubseteq A$. Then $\mathcal{T} \not\models D \sqsubseteq B$, for some conjunct B of F. By IH, $\models C \sqsubseteq D$ for the (unique) $C \in \mathsf{noimply}^0_{\mathcal{T},\Sigma}(B)$, and therefore $\models C \sqsubseteq D$ for some $C \in \mathsf{noimply}^0_{\mathcal{T},\Sigma}(A)$.

For the induction step, assume that **C2** has been shown for n. Let D be a Σ-concept with $\mathcal{T} \not\models D \sqsubseteq A$ and $\mathsf{depth}(D) = n+1$. Assume first that A is pseudo-primitive in $\mathcal{T}$. Then the atomic conjuncts of D are included in $\Sigma \setminus \mathsf{pre}^{\Sigma}_{\mathcal{T}}(A)$. So, from $C = \bigsqcap_{A' \in \Sigma \setminus \mathsf{pre}^{\Sigma}_{\mathcal{T}}(A)} A' \sqcap \mathsf{all}^{n+1}_{\Sigma}$ we obtain $\models C \sqsubseteq D$.

Now assume $A \equiv \exists r.B \in \mathcal{T}$. Let $C^{n+1}_{\mathcal{T},\Sigma}$ be the only element of $\mathsf{noimply}^{n+1}_{\mathcal{T},\Sigma}(A)$ and assume

$$D = \bigsqcap_{B \in Q_0} B \sqcap \bigsqcap_{(s,D') \in Q_1} \exists s.D'.$$

Then $Q_0 \subseteq \Sigma \setminus \mathsf{pre}^{\Sigma}_{\mathcal{T}}(A)$. Hence, $\models C^{n+1}_{\mathcal{T},\Sigma} \sqsubseteq \bigsqcap_{B \in Q_0} B$. Now consider a conjunct $\exists s.D'$ of D. There are two cases:

- $s \neq r$. Then, by construction, $\models C^{n+1}_{\mathcal{T},\Sigma} \sqsubseteq \exists s.D'$.
- $s = r$. It is enough to show that there exists $E \in \mathsf{noimply}^{n}_{\mathcal{T},\Sigma}(B)$ such that $\models E \sqsubseteq D'$. Suppose there does not exist such an E. Then, by IH, $\mathcal{T} \models D' \sqsubseteq B$. Hence, $\mathcal{T} \models \exists r.D' \sqsubseteq \exists r.B$ and we obtain $\mathcal{T} \models D \sqsubseteq A$, which is a contradiction.

The case in which A is conjunctive in $\mathcal{T}$ is straightforward and is left to the reader.

Corollary 1. *For all terminologies $\mathcal{T}'$ and $A \in \mathsf{N_C}$ the following are equivalent:*

1. *there exists a Σ-concept C such that $\mathcal{T} \not\models C \sqsubseteq A$ and $\mathcal{T}' \models C \sqsubseteq A$;*
2. *there exists $n \geq 0$ and $C \in \mathsf{noimply}^{n}_{\mathcal{T},\Sigma}(A)$ such that $\mathcal{T}' \models C \sqsubseteq A$.*

Proof. The direction from Point 2 to Point 1 follows immediately from **C1**. Conversely, assume that there exists a Σ-concept C such that $\mathcal{T}' \models C \sqsubseteq A$ and $\mathcal{T} \not\models C \sqsubseteq A$. By **C1** and **C2**, there exist n and $C' \in \mathsf{noimply}^{n}_{\mathcal{T},\Sigma}(A)$ with $\models C' \sqsubseteq C$ and $\mathcal{T} \not\models C' \sqsubseteq A$. Then $\mathcal{T}' \models C' \sqsubseteq A$.

In contrast to the formulation of Theorem 3, Corollary 1 does not provide us with a polynomial time algorithm. First, no upper bound on n is given and, second, the concepts in $\mathsf{noimply}^{n}_{\mathcal{T},\Sigma}(A)$ are of exponential size in n. Example 1 is easily extended so as to show that this is unavoidable: one can construct a terminology $\mathcal{T}$ and a sequence of terminologies $\mathcal{T}'_n$ such that in minimal implications in $\mathsf{Diff}_{\Sigma}(\mathcal{T}, \mathcal{T}')$ of the form $C_n \sqsubseteq A$ the concept C_n has at least depth n and is of size 2^n. However, Theorem 3 is now an immediate consequence of the following lemma and Corollary 1.

Lemma 5. *Let $\mathcal{T}'$ be a terminology such that $\mathsf{sig}(\mathcal{T}') \cap \Sigma^{\mathsf{fresh}} = \emptyset$ and $A \in \mathsf{sig}(\mathcal{T}) \cup \Sigma$. Then the following conditions are equivalent:*

1. *$\mathcal{T}' \cup \mathsf{Noimply}_{\mathcal{T},\Sigma} \models \xi \sqsubseteq A$, for some $\xi \in \mathsf{noimply}_{\mathcal{T},\Sigma}(A)$;*

2. $\mathcal{T}' \models C \sqsubseteq A$*, for some* $n \geq 0$ *and* $C \in \mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$*.*

Proof. Point 2 implies Point 1. For concept names A which are non-conjunctive in $\mathcal{T}$ this follows because $\mathsf{Noimply}_{\mathcal{T},\Sigma} \models \xi_A \sqsubseteq C$ for the only element C of $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$. The conjunctive case follows by induction.

Point 1 implies Point 2 is proved by a compactness argument. Intuitively, if $\mathcal{T}' \cup \bigcup_{n\geq 0} \mathsf{noimply}^n_{\mathcal{T},\Sigma}(A) \not\models A$, then $\mathcal{T}' \cup \mathsf{Noimply}_{\mathcal{T},\Sigma} \not\models \xi \sqsubseteq A$, for all $\xi \in \mathsf{noimply}_{\mathcal{T},\Sigma}(A)$. However, to prove this, one has to re-construct the concepts $\mathsf{noimply}^n_{\mathcal{T},\Sigma}(A)$; details of the proof are given in the technical report.

5 Practical algorithm and system

We have seen above that the sets

- $\mathsf{DiffR}_\Sigma(\mathcal{T},\mathcal{T}')$ consisting of all $A \in \Sigma$ such that there is a Σ-concept C with $\mathcal{T} \not\models C \sqsubseteq A$ and $\mathcal{T}' \models C \sqsubseteq A$, and
- $\mathsf{DiffL}_\Sigma(\mathcal{T},\mathcal{T}')$ consisting of all $A \in \Sigma$ such that there is a Σ-concept C with $\mathcal{T} \not\models A \sqsubseteq C$ and $\mathcal{T}' \models A \sqsubseteq C$

can be computed in polynomial time and can be regarded, by Lemma 2, as an informative approximation of the logical difference between $\mathcal{T}$ and $\mathcal{T}'$ w.r.t. Σ.

Computing both sets for large terminologies and signatures Σ using a direct implementation of the algorithm described above will fail: considering that state of the art description logic reasoners [2] take about 15 minutes to classify the SNOMED CT terminology [17], the reduction to reasoning given in Section 4 is impractical for large terminologies and signatures of reasonable size (the terminology $\mathsf{Noimply}_{\mathcal{T},\Sigma}$ contains huge conjunctions of Σ-concept names). We now discuss the implementation of the algorithms above in the system CEX for acyclic terminologies using a dynamic programming approach.

Let $\mathcal{T}$ and $\mathcal{T}'$ be acyclic terminologies and Σ a signature. For expositional reasons, we assume that $\Sigma \subseteq \mathsf{sig}(\mathcal{T}') \subseteq \mathsf{sig}(\mathcal{T})$. This is justified because we can add $A \sqsubseteq \top$ to $\mathcal{T}'$, for all $A \in \Sigma \setminus \mathsf{sig}(\mathcal{T}')$, and $A \sqsubseteq \top$ to $\mathcal{T}$, for all $A \in (\Sigma \cup \mathsf{sig}(\mathcal{T}')) \setminus \mathsf{sig}(\mathcal{T})$. We describe the algorithm computing DiffR_Σ, the rather straightforward algorithm computing DiffL_Σ is discussed in the technical report. We assume that $\mathcal{T}$ and $\mathcal{T}'$ are fully classified and the result of the classification is kept in a table, so, given two concept names A and B, it takes constant time to find out whether $\mathcal{T} \models A \sqsubseteq B$ (likewise, if $\mathcal{T}' \models A \sqsubseteq B$). Now the algorithm computing DiffR_Σ works by induction on concept definitions and marks, recursively, every $E \in \mathsf{sig}(\mathcal{T}')$, starting with pseudo-primitive ones, with members of $\Xi = \{\xi_A \mid A \in \mathsf{sig}(\mathcal{T})$ non-conjunctive in $\mathcal{T}\}$ in such a way that

(†) $E \in \mathsf{sig}(\mathcal{T}')$ is marked with ξ if, and only if, $\mathcal{T}' \cup \mathsf{Noimply}_{\mathcal{T},\Sigma} \not\models \xi \sqsubseteq E$.

Then $A \in \Sigma$ is *not* marked with $\xi \in \mathsf{noimply}_{\mathcal{T},\Sigma}(A)$ if, and only if, $\mathcal{T}' \cup \mathsf{Noimply}_{\mathcal{T},\Sigma} \models \xi \sqsubseteq A$. If this happens to be the case for some $\xi \in \mathsf{noimply}_{\mathcal{T},\Sigma}(A)$, then A is included in $\mathsf{DiffR}_\Sigma(\mathcal{T},\mathcal{T}')$ (Theorem 3).

In order to define the marking, set $\mathsf{pre}^\Sigma_{\mathcal{T}}(\xi_A) = \mathsf{pre}^\Sigma_{\mathcal{T}}(A)$, for $A \in \mathsf{sig}(\mathcal{T})$ non-conjunctive in $\mathcal{T}$. Now mark $E \in \mathsf{sig}(\mathcal{T}')$ as follows:

1. If E is pseudo-primitive in $\mathcal{T}'$, then it is marked with all $\xi \in \Xi$ such that $\mathsf{pre}^{\Sigma}_{\mathcal{T}'}(E) \subseteq \mathsf{pre}^{\Sigma}_{\mathcal{T}}(\xi)$;
2. If $E \equiv E_1 \sqcap \ldots \sqcap E_k \in \mathcal{T}'$, then it is marked with all $\xi \in \Xi$ such that at least one of $E_1, \ldots, E_k$ is marked with ξ;
3. If $E \equiv \exists r.E' \in \mathcal{T}'$ and
   (a) if $r \notin \Sigma$ or $\mathcal{T}' \cup \{\alpha\} \not\models (\bigsqcap_{A' \in \Sigma} A' \sqcap \mathsf{All}_{\Sigma}) \sqsubseteq E'$, then E is marked with all $\xi \in \Xi$ such that $\mathsf{pre}^{\Sigma}_{\mathcal{T}'}(E) \subseteq \mathsf{pre}^{\Sigma}_{\mathcal{T}}(\xi)$;
   (b) if $r \in \Sigma$ and $\mathcal{T}' \cup \{\alpha\} \models (\bigsqcap_{A' \in \Sigma} A' \sqcap \mathsf{All}_{\Sigma}) \sqsubseteq E'$, then E is marked with all $\xi_A \in \Xi$ such that
      - $A \equiv \exists r.A'$ in $\mathcal{T}$ and, for all $\xi' \in \mathsf{noimply}_{\mathcal{T},\Sigma}(A')$, E' is marked with ξ' and
      - $\mathsf{pre}^{\Sigma}_{\mathcal{T}'}(E) \subseteq \mathsf{pre}^{\Sigma}_{\mathcal{T}}(\xi_A)$.

Using Theorem 3 and Lemma 5, one can prove that the defined marking has property (†). While the condition $\mathcal{T}' \cup \{\alpha\} \models (\bigsqcap_{A' \in \Sigma} A' \sqcap \mathsf{All}_{\Sigma}) \sqsubseteq E$ can be checked directly, this requires operating concepts of large size for large Σ's. So, instead we use the following criterion: we may assume that $\mathcal{T}$ contains a definition $A \sqsubseteq \top$ such that $A \notin \mathsf{sig}(\mathcal{T}')$ and A does not occur elsewhere in $\mathcal{T}$. Then it follows from the definitions that $\mathcal{T}' \cup \{\alpha\} \models (\bigsqcap_{A' \in \Sigma} A' \sqcap \mathsf{All}_{\Sigma}) \sqsubseteq E$ if, and only if, E is not marked with ξ_A.

Let T and T' be the time taken to fully classify $\mathcal{T}$ and $\mathcal{T}'$, respectively. Then all $\mathcal{T}'$ concept names can be marked in $O(|\mathcal{T}| \times |\mathcal{T}'| \times |\Sigma| + T')$ time. Overall, checking Σ-entailment takes $O(|\mathcal{T}| \times |\mathcal{T}'| \times |\Sigma| + T + T')$ time and $O(|\mathcal{T}| \times |\mathcal{T}'| \times |\Sigma|)$ space. It should be noted that in our implementation this theoretical upper bound is often not reached due to the use of hash tables and structure sharing.

6 Experimental evaluation

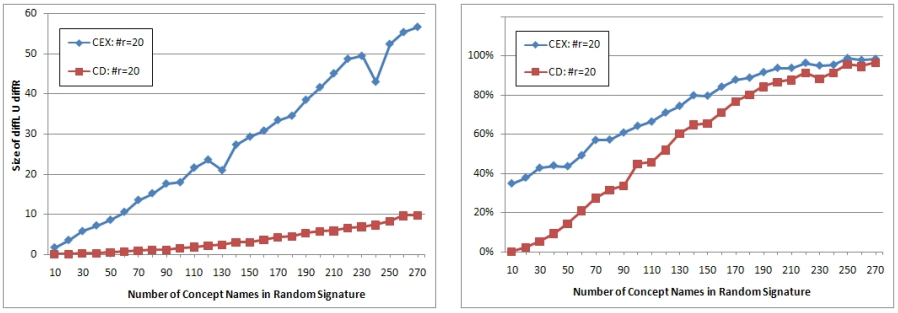
CEX (see http://www.csc.liv.ac.uk/∼konev/software/) is an OCaml program [4]. For the experiments, we use two versions of SNOMED CT: one dated 09 February 2005 (SM-05) and the other 30 December 2006 (SM-06) and having 379 691 and 389 472 axioms, respectively. As CEX currently accepts acyclic $\mathcal{EL}$-terminologies only, the role inclusions of SNOMED CT are not taken into account. The tests have been carried out on a standard PC: Intel® Core™ 2 CPU at 2.13 GHz and 3 GB of RAM.

Logical difference between SM-05 *and* SM-06. Table 1 shows the average sizes of the lists $\mathsf{DiffL}_{\Sigma}(\text{SM-05}, \text{SM-06})$ and $\mathsf{DiffR}_{\Sigma}(\text{SM-05}, \text{SM-06})$ for 20 randomly generated signatures $\Sigma \subseteq \mathsf{sig}(\text{SM-05}) \cap \mathsf{sig}(\text{SM-06})$ for each of the 12 possible signature sizes containing 100, 1 000, etc. concept names and 0, 20, or 40 role names.[4] The execution time and memory consumption of CEX when computing these lists vary from 477 to 596 seconds and from 1 393 to 1 496 MByte, respectively. The numbers show that there is a huge difference between SM-05 and SM-06. Also, adding a role name to the signature has a larger impact on the number of differences than adding a concept name.

[4] There are 50 role names in $\mathsf{sig}(\text{SM-05}) \cap \mathsf{sig}(\text{SM-06})$.

$\lvert\Sigma \cap \mathsf{N_C}\rvert$	$\lvert\Sigma \cap \mathsf{N_R}\rvert = 0$		$\lvert\Sigma \cap \mathsf{N_R}\rvert = 20$		$\lvert\Sigma \cap \mathsf{N_R}\rvert = 40$	
	$\lvert\mathsf{diffL}_\Sigma\rvert$	$\lvert\mathsf{diffR}_\Sigma\rvert$	$\lvert\mathsf{diffL}_\Sigma\rvert$	$\lvert\mathsf{diffR}_\Sigma\rvert$	$\lvert\mathsf{diffL}_\Sigma\rvert$	$\lvert\mathsf{diffR}_\Sigma\rvert$
100	0.10	0.10	0.90	0.15	2.95	0.20
1 000	2.35	2.15	15.55	2.95	28.85	3.75
10 000	155.35	125.35	257.35	136.20	514.10	209.90
100 000	11 795.90	4 108.60	12 954.45	4 358.30	14 942.55	6 823.60

Table 1. Computing logical difference with CEX: $\mathsf{Diff}_\Sigma(\text{SM-05}, \text{SM-06})$



(a) Number of differences (b) Proportion of detected differences

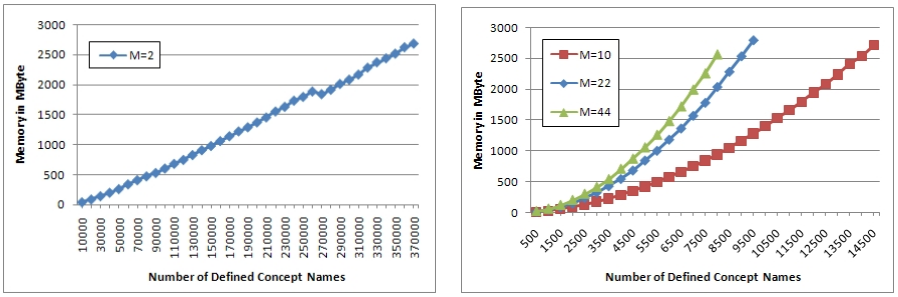
Fig. 4. Comparison of CEX and classification-based approach

Comparison with the classification approach. We compare the size of $\mathsf{DiffL}_\Sigma \cup \mathsf{DiffR}_\Sigma$ as computed by CEX with the number of concept names in Σ for which there is a difference in the class hierarchy restricted to Σ; i.e., the set of $A \in \Sigma$ such that there exists $B \in \Sigma$ such that $A \sqsubseteq B \in \mathsf{Diff}_\Sigma$ or $B \sqsubseteq A \in \mathsf{Diff}_\Sigma$. The experiments show how many of the differences between two terminologies detected by CEX can be extracted from a straightforward comparison of class hierarchies.

To facilitate the experiments, we use an empty terminology and an SM-05 fragment containing about 140 000 axioms. For every number between 10 and 270 with the step of 10, we generated 500 samples of a random signature containing this number of concepts and 20 roles. The results of the experiments are given in Figure 4. 4(a) shows that, for these signatures, the number of concept names CEX outputs is about five times larger than the number of concept names occurring in differences between the class hierarchies. In 4(b), we do not count the number of differences but analyse how often the two approaches detect differences at all. More precisely, we give the percentage of cases when CEX detects a difference between the two terminologies and when a difference is visible in the class hierarchies. For signatures larger than 200, both approaches almost always detect differences. But for smaller signatures there is again a significant gap between the two approaches.

Scalability. We demonstrated in the previous section that CEX is capable of finding the logical difference in two unmodified versions of SNOMED CT. In order to see how CEX's performance scales, we now test it on randomly generated acyclic

terminologies of various sizes. Each randomly generated terminology contains a certain number of defined- and primitive concept names and role names. The ratio between concept equations and concept inclusions is fixed, as is the ratio between existential restrictions and conjunctions. The random terminologies were generated for a varying number of defined concept names using the parameters of SM-05: 62 role names; the average number of conjuncts is 2.59; the equality-inclusion ratio is 0.102; and the exists-conjunction ratio is 0.652. For every chosen size, we generate a number of samples consisting of two random terminologies as described above. We apply CEX to find the logical difference of the two terminologies over their joint signature. Figure 5 shows the time and memory consumption of CEX on randomly generated terminologies of various sizes. In 5(a) the maximum length of conjunctions was fixed as two (M=2), and in 5(b) the number of conjuncts in each conjunction is randomly selected between two and M. It can be seen that the performance of CEX crucially depends on the length of conjunctions. In 5(b), the curves break off at the point where CEX runs out of memory. For instance, in the case M=22, this happens for terminologies with more than 9 500 defined concept names.



(a) Short conjunctions (b) Long conjunctions

Fig. 5. Memory consumption of CEX on randomly generated terminologies

7 Uniform Interpolation

Let $\mathcal{T}$ be a terminology and Σ a signature. A general TBox $\mathcal{T}_\Sigma$ is called a *uniform interpolant* for $\mathcal{T}$ w.r.t. Σ if $\mathsf{sig}(\mathcal{T}_\Sigma) \subseteq \Sigma$ and $\mathcal{T}_\Sigma$ and $\mathcal{T}$ are Σ-inseparable. The question whether uniform interpolants exist for every terminology $\mathcal{T}$[5] and signature Σ in a logic (i.e., whether the logic has *uniform interpolation*), has been investigated extensively in the literature, in particular in modal and intuitionistic logic [15, 18, 8]. For instance, modal logic K has uniform interpolation [18], but S4 does not [8]. Observe that, if a uniform interpolant $\mathcal{T}'_\Sigma$ of $\mathcal{T}'$ w.r.t. Σ

[5] In modal or intuitionistic logic $\mathcal{T}$ is, of course, a formula.

exists, then $\mathcal{T}$ Σ-entails $\mathcal{T}'$ if, and only if, $\mathcal{T} \models \mathcal{T}'_\Sigma$. Thus, the problem of deciding Σ-entailment is reduced to computing a uniform interpolant and standard deduction. Unfortunately, even for $\mathcal{EL}$-terminologies uniform interpolants do not always exist.

Lemma 6. *There exists an $\mathcal{EL}$-terminology $\mathcal{T}$ and a signature Σ such that there does not exist an uniform interpolant of $\mathcal{T}$ w.r.t. Σ.*

Proof. Let $\mathcal{T} = \{A_0 \sqsubseteq B, B \sqsubseteq A_1 \sqcap \exists r.B\}$ and $\Sigma = \{A_0, A_1, r\}$. Then a uniform interpolant $\mathcal{T}_\Sigma$ would have to axiomatise (using symbols from Σ only) the class of interpretations $\mathcal{I}$ satisfying the following condition: if $d_0 \in A_0^\mathcal{I}$, then there exists a sequence $d_0 r^\mathcal{I} d_1 r^\mathcal{I} d_2 r^\mathcal{I} \ldots$ with $d_i \in A_1^\mathcal{I}$ for all $i \geq 0$. It is not difficult to show that no such $\mathcal{T}_\Sigma$ exists (even in first-order logic).

On the other hand, uniform interpolants always exist for acyclic $\mathcal{EL}$-terminologies, but minimal uniform interpolants might contain exponentially many axioms.

Theorem 4. *Let $\mathcal{T}$ be an acyclic terminology and Σ a signature. Then there exists a uniform interpolant of $\mathcal{T}$ w.r.t. Σ. In the worst case, minimal uniform interpolants have exponentially many axioms.*

Proof. First, one can show that $\mathcal{T}_\Sigma = \mathcal{T}^l_\Sigma \cup \mathcal{T}^r_\Sigma$ is a uniform interpolant for $\mathcal{T}$ w.r.t. Σ if $\mathsf{sig}(\mathcal{T}_\Sigma) \subseteq \Sigma$ and

(a) $\mathcal{T} \models C \sqsubseteq A$ if, and only if, $\mathcal{T}^l_\Sigma \models C \sqsubseteq A$, for all Σ-concepts C and $A \in \Sigma$;
(b) $\mathcal{T} \models A \sqsubseteq D$ if, and only if, $\mathcal{T}^r_\Sigma \models A \sqsubseteq D$, for all Σ-concepts D and $A \in \Sigma$.

Due to space constraints we cannot describe the construction of $\mathcal{T}^l_\Sigma$ and $\mathcal{T}^r_\Sigma$ here, and refer the reader to the technical report. The following example shows that, in the worst case, minimal uniform interpolants require exponentially many axioms. Let

$$\mathcal{T} = \{A \equiv B_1 \sqcap \cdots \sqcap B_n\} \cup \{A_{ij} \sqsubseteq B_i \mid 1 \leq i, j \leq n\}.$$

and $\Sigma = \{A\} \cup \{A_{ij} \mid 1 \leq i, j \leq n\}$. Then

$$\mathcal{T}_\Sigma = \{A_{1j_1} \sqcap \cdots \sqcap A_{n,j_n} \sqsubseteq A \mid 1 \leq j_1, \ldots, j_n \leq n\}$$

is a uniform interpolant. It is easy to see that no uniform interpolant with fewer axioms exists. This example shows as well that one has to allow for general TBoxes when constructing uniform interpolants.

The results above show that, at least from a theoretical viewpoint, deciding Σ-entailment via uniform interpolants is less efficient than the approach discussed before. Still, uniform interpolants are useful for a number of applications, and it would be of interest to see whether this approach is viable for real-world terminologies.

8 Discussion

We have shown that computing the logical difference is tractable for $\mathcal{EL}$-terminologies and that this approach exhibits differences which are not visible in the class hierarchy. Our experiments with SNOMED CT show that the algorithm can be implemented in such a way that very large terminologies can be compared efficiently.

The following result shows that there is no straightforward way of extending these results to (even acyclic) terminologies in the basic Boolean description logic $\mathcal{ALC}$ (in which concepts can be constructed using, in addition, negation).

Theorem 5. *(1) Σ-entailment is* NEXPTIME*-hard for acyclic $\mathcal{ALC}$-terminologies. (2) Uniform interpolants do not always exist for acyclic $\mathcal{ALC}$-terminologies.*

Proof. Point (1) can be proved by a reduction of the NEXPTIME-hard problem of deciding conservative extensions in modal logic K [7], details are given in the technical report.

Point (2). We rewrite the terminology from Lemma 6. Let $\mathcal{T} = \{A \sqsubseteq (\neg A_0 \sqcup B) \sqcap (\neg B \sqcup (A_1 \sqcap \exists r.B))\}$ and $\Sigma = \{A, A_0, A_1, r\}$. It follows from the proof of Lemma 6 that there does not exist a general $\mathcal{ALC}$-TBox $\mathcal{T}_\Sigma^A$ axiomatising (using only the symbols from Σ) the class $\mathcal{S}$ of interpretations $\mathcal{I}$ satisfying the following conditions: $A^\mathcal{I} = \Delta^\mathcal{I}$ and if $d_0 \in A_0^\mathcal{I}$, then there exists a sequence $d_0 r^\mathcal{I} d_1 r^\mathcal{I} d_2 r^\mathcal{I} \ldots$ with $d_i \in A_1^\mathcal{I}$ for all $i \geq 0$. Now assume that there exists a uniform interpolant $\mathcal{T}_\Sigma$ of $\mathcal{T}$ w.r.t. Σ. Then $\mathcal{T}_\Sigma \cup \{A \equiv \top\}$ would be an axiomatisation of $\mathcal{S}$ and we have derived a contradiction.

Point (2) of Theorem 5 is slightly unexpected, because it shows that it is not possible to lift results from modal logic K (which has uniform interpolation) to acyclic $\mathcal{ALC}$-terminologies. Besides of considering extensions of our approach to languages with additional concept constructors, such as $\mathcal{ALC}$, directions for future research include terminologies with additional role boxes. SNOMED CT has an additional role box consisting of implications $r \sqsubseteq r'$, $r \circ s \sqsubseteq r$ (right-identities), and $s \circ r \sqsubseteq r$ (left-identities), where r, s, r' are role names. It is not difficult to extend the algorithm (and implementation) presented in this paper to terminologies containing implications of the first type, but it remains open whether Σ-entailment is still tractable for additional role boxes containing left- and right-identities.

Finally, for the system CEX to be useful in practice, the outputs DiffL_Σ and DiffR_Σ have to be expanded by suggesting, for $A \in \mathsf{DiffR}_\Sigma$, Σ-concepts C such that $C \sqsubseteq A \in \mathsf{Diff}_\Sigma$, and similarly for DiffL_Σ. Computing such C's is straightforward by unfolding the concept ξ_A relative to $\mathsf{Noimply}_{\mathcal{T},\Sigma}$. However, even this might not provide enough information, because for the user it could be difficult to find out which difference between the axioms of the two terminologies has caused a certain Σ-difference. Thus, as a second step one might consider pinpointing algorithms *explaining* from which axioms of a terminology a counterexample $C \sqsubseteq A$ is derivable [3].

Acknowledgements. The authors were supported by EPSRC grant EP/E065279/1.

References

1. F. Baader. Terminological cycles in a description logic with existential restrictions. In *Proceedings of IJCAI'03*, pp. 325–330. Morgan Kaufmann, 2003. Long version available as LTCS Report 02-02.
2. F. Baader, C. Lutz, and B. Suntisrivaraporn. CEL—a polynomial-time reasoner for life science ontologies. In *Proceedings of IJCAR'06*, vol. 4130 of *LNAI*, pp. 287–291. Springer-Verlag, 2006.
3. F. Baader, R. Peñaloza, and B. Suntisrivaraporn. Pinpointing in the description logic $\mathcal{EL}^+$. In *Proceedings of KI'07*, vol. 4667 of *LNAI*, pp. 52–67. Springer, 2007.
4. The Caml team. `http://caml.inria.fr/contact.en.html`.
5. A. Flögel, H. K. Büning, and T. Lettmann. On the restricted equivalence of subclasses of propositional logic. *ITA*, 27(4):327–340, 1993.
6. S. Ghilardi, C. Lutz, and F. Wolter. Did I damage my ontology? a case for conservative extensions in description logics. In *Proceedings of KR'06*, pp. 187–197. AAAI Press, 2006.
7. S. Ghilardi, C. Lutz, F. Wolter, and M. Zakharyaschev. Conservative extensions in modal logics. In *Proceedings of AiML-6*, pp. 187–207. College Publications, 2006.
8. S. Ghilardi and M. Zawadowski. Undefinability of propositional quantifiers in the modal system S4. *Studia Logica*, 55(2):259–271, 1995.
9. B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler. Just the right amount: extracting modules from ontologies. In *Proceedings of WWW'07*, pp. 717–726. ACM, 2007.
10. M. Hofmann. Proof-theoretic approach to description logic. In *Proceedings of LICS'05*, pp. 229–237. IEEE Computer Society, 2005.
11. B. Konev, D. Walther, and F. Wolter. The logical difference problem for description logic terminologies. `http://www.csc.liv.ac.uk/~frank/publ/publ.html`, manuscript, 2008.
12. C. Lutz, D. Walther, and F. Wolter. Conservative extensions in expressive description logics. In *Proceedings of IJCAI'07*, pp. 453–458. AAAI Press, 2007.
13. C. Lutz and F. Wolter. Conservative extensions in the lightweight description logic $\mathcal{EL}$. In *Proceedings of CADE'07*, vol. 4603 of *LNCS*, pp. 84–99. Springer, 2007.
14. N. F. Noy and M. Musen. Promptdiff: A fixed-point algorithm for comparing ontology versions. In *Proceedings of AAAI'02*, pp. 744–750. AAAI Press, 2002.
15. A. Pitts. On an interpretation of second-order quantification in first-order intuitionistic propositional logic. *Journal of Symbolic Logic*, 57(1):33–52, 1992.
16. V. Sofronie-Stokkermans. Interpolation in local theory extensions. In *IJCAR'06*, pp. 235–250, 2006.
17. K. Spackman. Managing clinical terminology hierarchies using algorithmic calculation of subsumption: Experience with SNOMED-RT. *JAMIA*, 2000. Fall Symposium Special Issue.
18. A. Visser. Uniform interpolation and layered bisimulation. In *Gödel'96 (Brno, 1996)*, vol. 6 of *Lecture Notes Logic*, pp. 139–164. Springer, 1996.