Properties of Conservative Extensions and Modules

C. Lutz, U. Sattler, and F. Wolter

July 31, 2008

Today: Modularity for Lightweight Description Logics

Currently, two families of lightweight DLs are investigated and applied:

- DL-Lite: for conceptual modeling, data integration, querying instance data using background theories.
- $\mathcal{EL}$: many large medical and biologial ontologies are given in $\mathcal{EL}$.

We consider $\mathcal{EL}$ (but modularity in DL-Lite has been investigated as well).

Description Logic $\mathcal{EL}$

$\mathcal{EL}$ is a fragment of $\mathcal{ALC}$. Concept language of $\mathcal{EL}$:

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

TBox T is a finite set of concept inclusions $C \sqsubseteq D$.

Reasoning services:

- Satisfiability: every $\mathcal{EL}$-concept C is satisfiable w.r.t. any $\mathcal{EL}$-TBox.
- Subsumption: given C, D, T, does $T \models C \sqsubseteq D$? This problem is decidable in polynomial time.

Modularity reasoning for $\mathcal{EL}$

- Deciding whether two $\mathcal{EL}$-TBoxes are Σ-inseparable w.r.t. $\mathcal{EL}$ is ExpTime-complete.
- For $\mathcal{EL}$-TBoxes, Σ-inseparability w.r.t. SO is undecidable.
- For $\mathcal{EL}$-TBoxes, even $T \equiv_{\Sigma}^{SO} \emptyset$, (equivalently, whether

$$\{M_{|\Sigma} \mid M \models T\} = \text{ class of all } \Sigma\text{-models})$$

  is undecidable.
- $\mathcal{EL}$ has interpolation, but $(\mathcal{EL}, \mathcal{EL})$ is not robust under replacement.

Today, we consider $\mathcal{EL}$-TBoxes of a particular form.

$\mathcal{EL}$-terminologies

Definition

An $\mathcal{EL}$-TBox T is a $\mathcal{EL}$-terminology if

- every axiom is of the form $A \equiv C$, where A is a concept name;
- no concept name A occurs more than once on the left hand side of an axiom.

A $\mathcal{EL}$-terminology T is acyclic if no concept name refers to itself along definitions:

- let $A \prec_T X$ if there exists an axiom $A \equiv C$ in T such that X occurs in C.

Then T is acyclic iff $\prec_T$ is acyclic (equivalently $\prec_T^+$ is irreflexive).

In a TBox T, we write $A \sqsubseteq C$ instead of $A \equiv X \sqcap C$, if this is the only occurrence of X (and X not in any relevant signature Σ).

Plan for $\mathcal{EL}$-terminologies

- Deciding '$T \equiv_{\Sigma}^{SO} \emptyset$' in polynomial time (Uli: then T is safe!).
- Extracting modules.
- Logical difference: comparing versions of ontologies.
- Uniform interpolation (forgetting).

Deciding '$T \equiv_{\Sigma}^{SO} \emptyset$'

Theorem

The following problem can be solved in polynomial time: given an acyclic $\mathcal{EL}$-terminology T, decide whether

$$T \equiv_{\Sigma}^{SO} \emptyset.$$

For the proof, we distinguish two types of dependencies between Σ-symbols:

- in T, the 'definition' of a Σ-symbol uses another Σ-symbols;
- in T, two Σ-symbols are 'defined' using common non-Σ-symbols.

Syntactic Σ-dependencies

Let T be an acyclic $\mathcal{EL}$-terminology.

- T contains a syntactic Σ-dependency if there exist $A, X \in \Sigma$ such that $A \prec_T^+ X$.

Theorem

If an acyclic $\mathcal{EL}$-terminology T contains a syntactic Σ-dependency, then $T \not\equiv_\Sigma^{SO} \emptyset$.

Proof. Suppose T contains a syntactic Σ-dependency $A \prec_\Sigma^+ X$. Take a Σ-model M with $A^M = \Delta^M$ and $X^M = \emptyset$. Then M can't be expanded to a model of T.

Decomposing an acyclic $\mathcal{EL}$-terminology

Consider an acyclic $\mathcal{EL}$-terminology T. Take partition

$$T = T_\Sigma \cup T',$$

where

$$T_\Sigma = \{A \equiv C \mid A \in \Sigma \text{ or } \exists B \in \Sigma \;\; B \prec_T^+ A\}$$

Theorem

If $M \models T_\Sigma$, then there exists M' with $M'_{|\Sigma} = M_{|\Sigma}$ such that $M' \models T$.

Proof. Expand M inductively by setting $A^{M'} := C^{M'}$ for $A \equiv C \in T'$.

Checking indirect Σ-dependencies

Theorem

Let T be an acyclic $\mathcal{EL}$-terminology without syntactic Σ-dependencies. Then the following conditions are equivalent:

- $T \equiv_{\Sigma}^{SO} \emptyset$;
- *Every one point Σ-model can be expanded to a model of T_{Σ}.*

Point 2 implies Point 1. Let M be a Σ-model. As T_{Σ} contains no Σ-roles, we may assume that Σ contains no roles. For each $d \in M$, let $M'_{\{d\}} \models T_{\Sigma}$ be an expansion of $M_{\{d\}}$. Then

$$M' = \bigcup_{d \in M} M'_{\{d\}} \models T_{\Sigma}$$

and M' is an expansion of M.

Polytime algorithm for $T \equiv_{\Sigma}^{SO} \emptyset$

To decide whether $T \equiv_{\Sigma}^{SO} \emptyset$, check

- T contains no syntactic Σ-dependencies;
- every one point Σ-model can be expanded to a model of T_{Σ}.

Point 2 holds iff

For all $A \in \Sigma$,

$$\{X \mid A \prec_T^+ X\} \not\subseteq \{X \mid \exists B \in \Sigma \setminus \{A\} \;\; B \prec_T^+ X\}.$$

Observation: For acyclic $\mathcal{ALC}$-terminologies without syntactic Σ-dependencies, one can decide $T \equiv_{\Sigma}^{SO} \emptyset$ by considering one point-models (then Π_2^p-complete).

Module extraction

Given T and Σ, we use this decision procedure to extract from T the smallest $M \subseteq T$ such that

$$T \setminus M \equiv^{SO}_{\Sigma \cup \mathsf{sig}(M)} \emptyset.$$

Equivalently,

$$M \equiv^{SO}_{\Sigma \cup \mathsf{sig}(M)} T.$$

Uli: then $T \setminus M$ is safe for $\Sigma \cup \mathsf{sig}(M)$.

Module extraction algorithm

Input: acyclic $\mathcal{EL}$-terminology T and signature Σ.

Output: Smallest $M \subseteq T$ such that $T \setminus M \equiv_{\Sigma \cup \mathsf{sig}(M)} \emptyset$.

Initialise: $M = \emptyset$, $\Sigma' = \Sigma$. Apply Rules 1 and 2 exhaustively, prefering Rule 1.

- if $A \in \Sigma'$, $A \equiv C \in T \setminus M$, and exists $X \in \Sigma'$ with $A \prec^{+}_{T \setminus M} X$,

$$M := M \cup \{A \equiv C\}, \quad \Sigma' := \Sigma' \cup \mathsf{sig}(C).$$

- if $A \in \Sigma'$, $A \equiv C \in T \setminus M$, and

$$\{X \mid A \prec^{+}_{T \setminus M} X\} \subseteq \{X \mid \exists B \in \Sigma' \setminus \{A\} \;\; B \prec^{+}_{T \setminus M} X\},$$

then set

$$M := M \cup \{A \equiv C\}, \quad \Sigma' := \Sigma' \cup \mathsf{sig}(C).$$

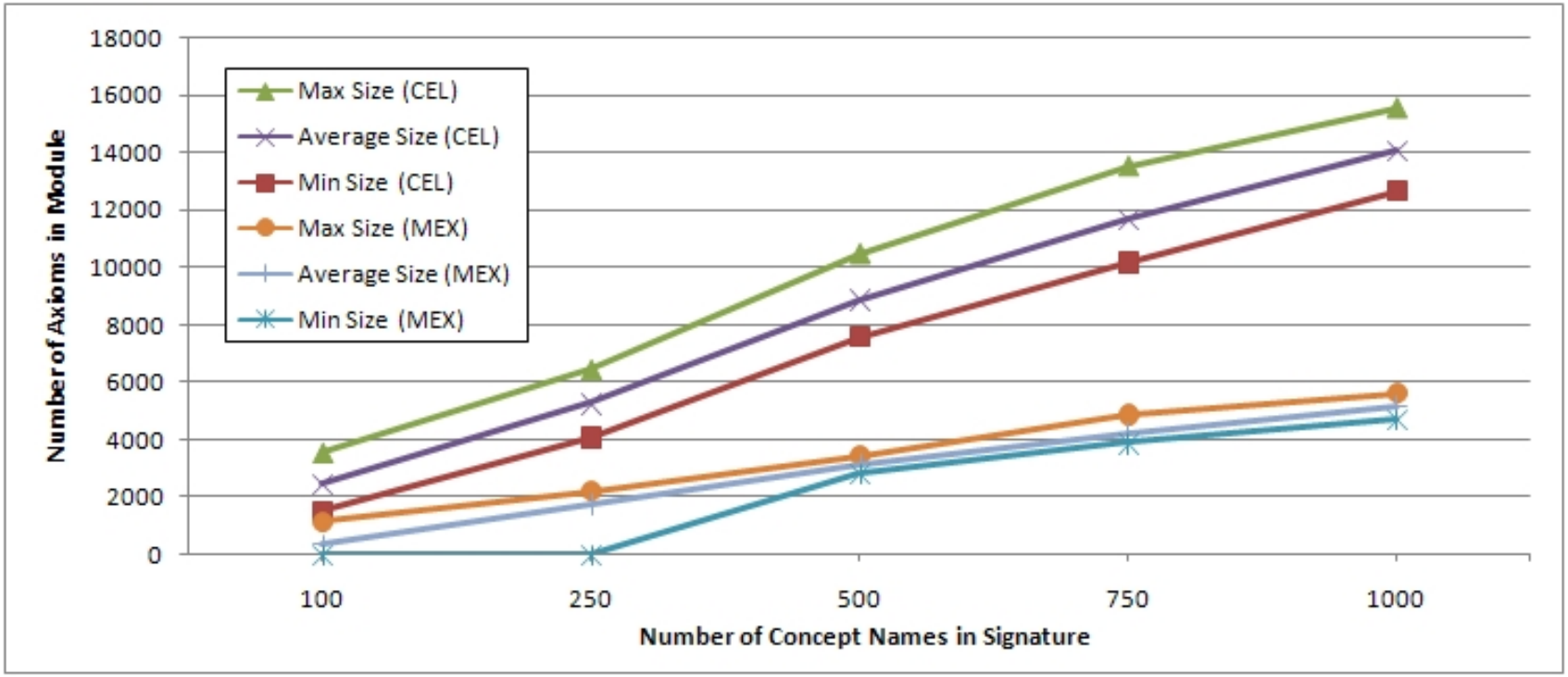
Experiment with SNOMED CT

SNOMED CT:

- Systematised Nomenclature of Medicine (Clinical Terms).
- $\sim 400,000$ terms
- used in health care etc. in the US, UK, Australia etc.
- an acyclic $\mathcal{EL}$-terminology ($+$ role box):

Experiment: Extraction of modules from SNOMED CT

- We use implementation MEX of the algorithm above.
- Σ — randomly selected from SNOMED CT.
- 1000 samples for each signature size
- with role box (under simplifying assumptions)



Logical Difference: motivation

Problem: given two versions T_1 and T_2 of an ontology and a signature Σ, compute "the difference" between T_1 and T_2 observable in Σ in a query language $\mathcal{QL}$.

Lot's of tools to compute the difference between versions of texts and program code! But this is not of much help for ontologies/logical theories:

- $T_1 = \{A \sqsubseteq B_1 \sqcap B_2\}$
- $T_2 = \{A \sqsubseteq B_1, A \sqsubseteq B_2\}$.
- $\Sigma = \{A, B_1, B_2\}$.

Then $T_1 \neq T_2$, but T_1 and T_2 are equivalent (so Σ-inseparable w.r.t. SO).

Logical Difference

T_1 and T_2 ontologies, $\mathcal{QL}$ a query language, Σ a signature.
The logical Σ-difference between T_1 and T_2 w.r.t. $\mathcal{QL}$ is defined as

$$\mathrm{Diff}_{\Sigma}^{\mathcal{QL}}(T_1, T_2) \cup \mathrm{Diff}_{\Sigma}^{\mathcal{QL}}(T_2, T_1),$$

where

- $\mathrm{Diff}_{\Sigma}^{\mathcal{QL}}(T_1, T_2) = \{\varphi \in \mathcal{QL}_{\Sigma} \mid T_1 \models \varphi, T_2 \not\models \varphi\}$.
- $\mathrm{Diff}_{\Sigma}^{\mathcal{QL}}(T_2, T_1) = \{\varphi \in \mathcal{QL}_{\Sigma} \mid T_2 \models \varphi, T_1 \not\models \varphi\}$.

Observation: $\mathrm{Diff}_{\Sigma}^{\mathcal{QL}}(T_1, T_2) \cup \mathrm{Diff}_{\Sigma}^{\mathcal{QL}}(T_2, T_1) = \emptyset$ iff T_1 and T_2 are Σ-inseparable w.r.t. $\mathcal{QL}$.

Problem: How to present $\mathrm{Diff}_{\Sigma}^{\mathcal{QL}}(T_1, T_2)$ if it is non-empty?

Σ-difference for $\mathcal{EL}$-terminologies

Take query language $\mathcal{EL}$ consisting of $C \sqsubseteq D$, where C, D are $\mathcal{EL}$-concepts. We set

$$\mathrm{Diff}_\Sigma(T_1, T_2) = \mathrm{Diff}_\Sigma^{\mathcal{EL}}(T_1, T_2).$$

Example of 'large' smallest elements in $\mathrm{Diff}_\Sigma(T_1, T_2)$:

- $T_2 = \emptyset$;
- $T_1 = \{A' \sqsubseteq B_0, A \equiv B_n\} \cup \{B_{i+1} \equiv \exists r.B_i \sqcap \exists s.B_i \mid i < n\}$;
- $\Sigma = \{A', A, r, s\}$.

For the minimal $C \sqsubseteq A \in \mathrm{Diff}_\Sigma(T_1, T_2)$ we have $|C| = 2^n$.

Σ-difference for $\mathcal{EL}$-terminologies

Theorem

If $(C \sqsubseteq D) \in \mathrm{Diff}_\Sigma(T_1, T_2)$ *then either*

- $(A \sqsubseteq D_0) \in \mathrm{Diff}_\Sigma(T_1, T_2)$ *or*
- $(C_0 \sqsubseteq A) \in \mathrm{Diff}_\Sigma(T_1, T_2)$,

where A *is a concept name and*

A, C_0 — *subconcepts of* C;

D_0, A — *subconcepts of* D, *resp.*

In propositional $\mathcal{EL}$: if $C \sqsubseteq A_1 \sqcap A_2 \in \mathrm{Diff}_\Sigma(T_1, T_2)$, then

- $C \sqsubseteq A_1 \in \mathrm{Diff}_\Sigma(T_1, T_2)$ or
- $C \sqsubseteq A_2 \in \mathrm{Diff}_\Sigma(T_1, T_2)$.

Compact representation of $\mathsf{Diff}_\Sigma(T_1, T_2)$

Let

- $\mathsf{diffL}_\Sigma(T_1, T_2) =$
$$\left\{ A \in \Sigma \;\middle|\; \begin{array}{l} \text{there is a } \Sigma\text{-concept } C \text{ in } \mathcal{EL} \text{ s.t.} \\ T_1 \models A \sqsubseteq C \text{ and } T_2 \not\models A \sqsubseteq C \end{array} \right\}$$

- $\mathsf{diffR}_\Sigma(T_1, T_2) =$
$$\left\{ A \in \Sigma \;\middle|\; \begin{array}{l} \text{there is a } \Sigma\text{-concept } C \text{ in } \mathcal{EL} \text{ s.t.} \\ T_1 \models C \sqsubseteq A \text{ and } T_2 \not\models C \sqsubseteq A \end{array} \right\}$$

$\mathsf{diffL}_\Sigma(T_1, T_2)$ and $\mathsf{diffR}_\Sigma(T_1, T_2)$ provide a list of concept names in Σ about which T_1 "says more" than T_2.

Σ-difference between $\mathcal{EL}$-terminologies

Theorem

Let T_1 and T_2 be $\mathcal{EL}$-terminologies and Σ a signature. Then

- $\mathsf{diffL}_\Sigma(T_1, T_2)$ *and*
- $\mathsf{diffR}_\Sigma(T_1, T_2)$

can be computed in polynomial time. In particular, Σ-inseparability w.r.t. $\mathcal{EL}$ is tractable.

Proof idea for $\mathrm{DiffR}_\Sigma(T_1, T_2)$

Consider T_2 and $A \in \Sigma$. Compute set

$$\mathrm{pre}_\Sigma^{T_2}(A)$$

of most specific Σ-concepts C such that $T_2 \not\models C \sqsubseteq A$.
More precisely, $\mathrm{pre}_\Sigma^{T_2}(A)$ is a minimal set of Σ-concepts such that

- $T_2 \not\models C \sqsubseteq A$, for all $C \in \mathrm{pre}_\Sigma^{T_2}(A)$;
- if $T_2 \not\models D \sqsubseteq A$ and D a Σ-concept, then $\models C \sqsubseteq D$, for some $C \in \mathrm{pre}_\Sigma^{T_2}(A)$.

Suppose $\mathrm{pre}_\Sigma^{T_2}(A)$ has been computed. Then the following are equivalent:

- $A \in \mathrm{DiffR}_\Sigma(T_1, T_2)$;
- $T_1 \models C \sqsubseteq A$, for some $C \in \mathrm{pre}_\Sigma^{T_2}(A)$.

Computing $\mathrm{pre}_{\Sigma}^{T_2}(A)$ (propositional case)

Observation: In general, $\mathrm{pre}_{\Sigma}^{T_2}(A)$ is infinite and contains concepts of exponential size. However, polynomial representation possible using a TBox.

Consider acyclic propositional case:

A is primitive in T_2 if it has no occurence of the form $A \equiv D$ in T_2.

Then $\mathrm{pre}_{\Sigma}^{T_2}(A)$ can be defined inductively:

- For A primitive: $\mathrm{pre}_{\Sigma}(A) = \{\bigsqcap_{B \in \Sigma, T_2 \not\models B \sqsubseteq A} B\}$;
- For $A \equiv B_1 \sqcap \cdots \sqcap B_n \in T_2$: $\mathrm{pre}_{\Sigma}(A) = \bigcup_{i \leq n} \mathrm{pre}_{\Sigma}(B_i)$.

CEX applied to SNOMED CT

- CEX: implementation of tractable algorithm computing $\mathsf{DiffL}_\Sigma(T_1, T_2)$ and $\mathsf{DiffR}_\Sigma(T_1, T_2)$ for acyclic $\mathcal{EL}$-terminologies.
- Two versions of SNOMED CT: SNOMED CT'05 vs SNOMED CT'06

SNOMED CT'05 vs SNOMED CT'06

- Σ — randomly selected from $\mathrm{sig}(\mathit{SNOMED\ CT'05}) \cap \mathrm{sig}(\mathit{SNOMED\ CT'06})$
- 20 samples for every signature size

Size of Σ	CEX: diff(SNOMED CT'05,SNOMED CT'06)			
	Time (Sec.)	Memory (MByte)	$\lvert \mathrm{diffL}_\Sigma \rvert$	$\lvert \mathrm{diffR}_\Sigma \rvert$
100	513.1	1 393.7	0.0	0.0
1 000	512.4	1 394.6	2.5	2.5
10 000	517.7	1 424.3	183.2	122.0
100 000	559.8	1 473.2	11 322.1	4 108.5

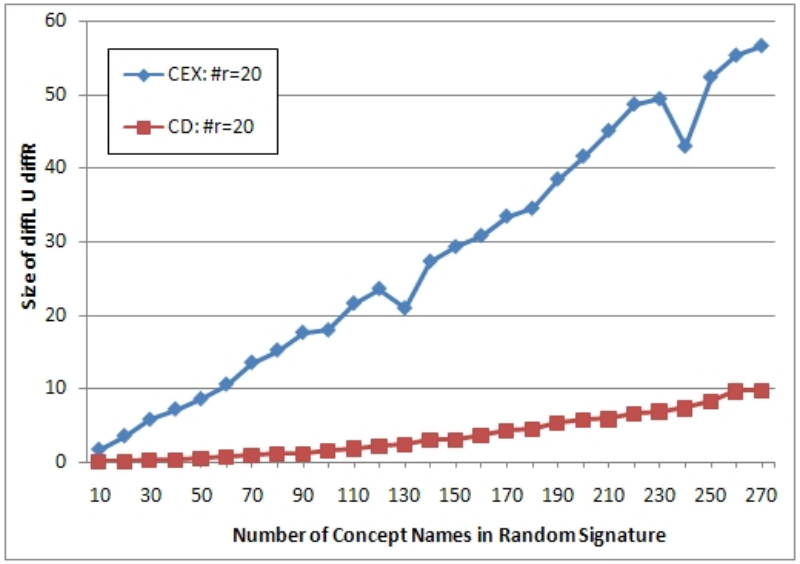
- Rolebox ignored

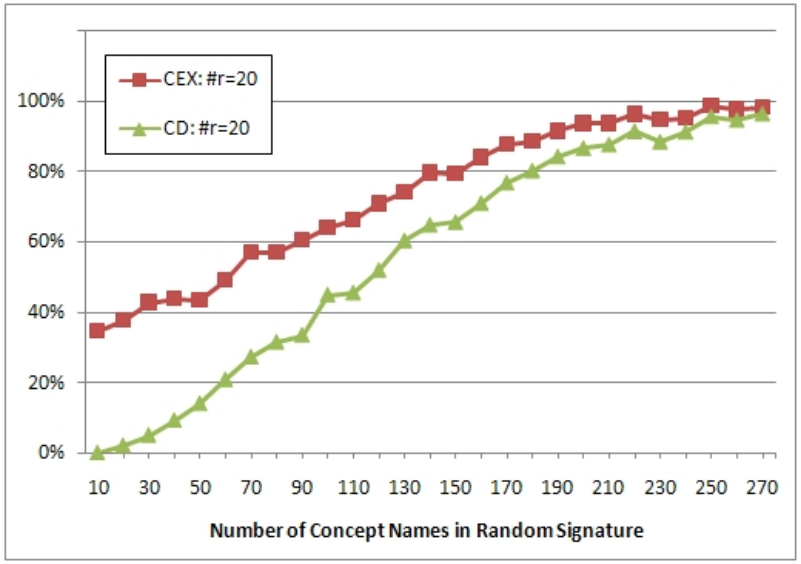
Comparison on the Joint Signature

- diff(SNOMED CT'05,SNOMED CT'06) on
  $\Sigma = \text{sig}(\mathit{SNOMED\ CT'}05) \cap \text{sig}(\mathit{SNOMED\ CT'}06)$
  - 689 seconds
  - $|\text{diffL}_\Sigma| + |\text{diffR}_\Sigma| = 162010$
  - Class hierarchy comparison misses 32475 of them

Comparing with $\emptyset$

- Combined $\mathrm{diffL}_\Sigma(\emptyset, S)$ and $\mathrm{diffR}_\Sigma(\emptyset, S)$
  - S — a part of SNOMED CT'05 containing $\sim 140,000$ axioms
  - Σ — randomly selected from M
  - 500 samples for each signature size
- Difference in class hierarchy





CEX on MEX

Instead of computing $\mathsf{diffL}_\Sigma(T_1, T_2) \cup \mathsf{diffR}_\Sigma(T_1, T_2)$ directly,

- first extract minimal Σ-modules T_1' and T_2' from T_1 and T_2, respectively,
- then compute $\mathsf{diffL}_\Sigma(T_1', T_2') \cup \mathsf{diffR}_\Sigma(T_1', T_2')$.

Size of Σ	CEX: diff(SNOMED CT'05,SNOMED CT'06)				CEX: diff(Mod'05,Mod'06)	
	Time (Sec.)	Memory (MByte)	$\lvert\mathsf{diffL}_\Sigma\rvert$	$\lvert\mathsf{diffR}_\Sigma\rvert$	Time (Sec.)	Memory (MByte)
100	513.1	1 393.7	0.0	0.0	3.66	116.5
1 000	512.4	1 394.6	2.5	2.5	4.46	122.5
10 000	517.7	1 424.3	183.2	122.0	22.29	126.5
100 000	559.8	1 473.2	11 322.1	4 108.5	189.98	615.8

Uniform Interpolation

Let T be a $\mathcal{EL}$ TBox and Σ a signature. A TBox T' is called a uniform interpolant of T w.r.t. Σ if the following holds:

- $\mathrm{sig}(T') \subseteq \Sigma$;
- $T \equiv_{\Sigma}^{\mathcal{EL}} T'$.

Theorem

Let T_1', T_2' be uniform interpolants of T_1 and T_2 w.r.t. Σ. The following conditions are equivalent:

- $T_1 \equiv_{\Sigma}^{\mathcal{EL}} T_2$;
- *T_1' and T_2' are logically equivalent.*

$\mathcal{EL}$-terminologies

Theorem

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

Proof. Let

$$T = \{A_0 \sqsubseteq B, B \sqsubseteq A_1 \sqcap \exists r.B\}, \quad \Sigma = \{A_0, A_1, r\}.$$

A uniform interpolant T_Σ would have to finitely axiomatise the class of models M satisfying:

- if $d_0 \in A_0^M$, then there exists a sequence $d_0 r^M d_1 r^M d_2 r^M \ldots$ with $d_i \in A_1^M$ for all $i \geq 0$.

Such a T_Σ does not exist (even in first-order logic).

Acyclic $\mathcal{EL}$-terminologies

Theorem

For acyclic $\mathcal{EL}$-terminologies, uniform interpolants always exist. In the worst case, exponentially many axioms are required.

Proof of second part. Let

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

and

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

Then

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

is a minimal uniform interpolant.