Mathematical Logic for Life Science Ontologies

Frank Wolter

Based on joint work with S. Ghilardi, B. Konev, R. Kontchakov, C. Lutz, D. Walther, M. Zakharyaschev

August 2, 2009

Plan

- Large-scale ontologies
- Conservative extensions/uniform interpolation
- Description logics $\mathcal{ALC}$ and $\mathcal{EL}$
- Conservative extensions/uniform interpolation in $\mathcal{ALC}$ and $\mathcal{EL}$.
- Experiments with SNOMED CT.

Large-scale terminologies/ontologies

- Life sciences, healthcare, and other knowledge intensive areas depend on having a common language for gathering and sharing knowledge.
- Common language is provided by reference terminologies.
- Reference terminologies often have more than 100 000 terms.
- Trend towards axiomatizing reference terminologies in weak fragments of first-order logic (typically description logics).
- Examples:
  - SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms),
  - NCI (National Cancer Institute Thesaurus),
  - FMA (Foundational Model of Anatomy),
  - GALEN (Medical Ontology), etc.

Reference terminology snippet

$$
\begin{array}{rcl}
\text{Cystic\_Fibrosis} & \sqsubseteq & \text{Fibrosis} \sqcap \exists \text{located\_In.Pancreas} \sqcap \exists \text{has\_Origin.Genetic\_Origin} \\
\text{Genetic\_Fibrosis} & \equiv & \text{Fibrosis} \sqcap \exists \text{has\_Origin.Genetic\_Origin} \\
\text{Genetic\_Fibrosis} & \sqsupseteq & \text{Fibrosis} \sqcap \exists \text{located\_In.Pancreas} \\
\text{Genetic\_Fibrosis} & \sqsubseteq & \text{Genetic\_Disorder} \\
\text{DEFBI\_Gene} & \sqsubseteq & \text{Immuno\_Protein\_Gene} \sqcap \exists \text{associated\_With.Cystic\_Fibrosis}
\end{array}
$$

In first-order logic syntax, for example:

$$\forall x.(\text{Genetic\_Fibrosis}(x) \leftrightarrow (\text{Fibrosis}(x) \wedge \exists y.\text{has\_Origin}(x, y) \wedge \text{Genetic\_Origin}(y)))$$

Example: SNOMED CT

- Comprehensive healthcare terminology consisting of 400 000 terms and approximately the same number of axioms.
- Property rights owned by not-for-profit organisation IHSTDO (International Health terminology Standards Development Organisation).
- IHSTDO made currently of nine nations (free in 49 developing countries).
- Aim: enabling clinicians, researchers and patients to share and exchange healthcare and clinical knowledge worldwide.
- Conference KR-MED-2008 devoted to SNOMED CT attracted more than 100 researchers.

SNOMED CT Snippet

$$
\begin{aligned}
\mathsf{EntireFemur} &\sqsubseteq \mathsf{StructureOfFemur} \\
\mathsf{FemurPart} &\sqsubseteq \mathsf{StructureOfFemur} \sqcap \\
&\quad \exists \mathsf{part\_of}.\mathsf{EntireFemur} \\
\mathsf{BoneStructureOfDistalFemur} &\sqsubseteq \mathsf{FemurPart} \\
\mathsf{EntireDistalFemur} &\sqsubseteq \mathsf{BoneStructureOfDistalFemur} \\
\mathsf{DistalFemurPart} &\sqsubseteq \mathsf{BoneStructureOfDistalFemur} \sqcap \\
&\quad \exists \mathsf{part\_of}.\mathsf{EntireDistalFemur} \\
\mathsf{StructureofDistalEpiphysisOfFemur} &\sqsubseteq \mathsf{DistalFemurPart} \\
\mathsf{EntireDistalEpiphysisOfFemur} &\sqsubseteq \mathsf{StructureOfDistalEpiphysisOfFemur}
\end{aligned}
$$

How is SNOMED CT used?

The concept hierarchy induced by a logical theory T is defined as

$$\{A \sqsubseteq B \mid A, B \text{ unary predicates in } T, T \models \forall x.A(x) \rightarrow B(x)\}.$$

Example:

$$\begin{array}{c} \text{EntireDistalEpiphysisOfFemur} \\ \sqsubseteq \\ \text{StructureOfDistalEpiphysisOfFemur} \\ \sqsubseteq \\ \text{DistalFemurPart} \\ \sqsubseteq \\ \text{BoneStructureOfDistalFemur} \\ \sqsubseteq \\ \text{FemurPart} \end{array}$$

Concept Hierarchy

Standard applications based on concept hierarchy:

- SNOMED CT is used to produce a hierarchy of medical terms. Each term is annotated with a numerical code and an axiom defining its meaning.
- This hierarchy is used by physicians to
  - generate,
  - process,
  - store,
  - share

  electronic medical records (EMRs) containing diagnoses, treatments, medication, lab records, etc.

Query Electronic Medical Records using SNOMED CT

Assume EMRs are given as a set $\mathcal{A}$ of ground facts

$$R(a, b), \quad P(c, d), \quad C(e), \quad \text{etc}$$

Query $\mathcal{A}$ using SNOMED CT; i.e., retrieve all $\vec{c}$ such that

$$(\text{SNOMED CT}, \mathcal{A}) \models \varphi(\vec{c}),$$

where φ is, e.g., a conjunctive query (FO-formula constructed from atoms using $\wedge$ and $\exists$).

Developing and using SNOMED CT

Many different versions

- When extending SNOMED CT, typically two terminologists axiomatize an extension, the outcome is discussed and the "official" extension is agreed upon.
- Different versions because of different standards in different countries.

Mistakes occur

- SNOMED CT $\models$ Amputation_of_arm $\sqsubseteq$ Amputation_of_hand

Small Σ enough

- Many applications use a very small subset of the signature of SNOMED CT only.

Support from logic for developing and using ontologies

Let Σ be a signature (a subject matter). We are interested in

- versioning: check whether T_1 and T_2 'say the same about' Σ;
- module extraction: compute minimal $M \subseteq T$ such that M and T 'say the same about' Σ.
- uniform interpolation: compute finite T_Σ such that T_Σ uses Σ only and T and T_Σ 'say the same about' Σ.

Formalise

'T_1 and T_2 say the same about Σ'.

Two formalisations of 'T_1 and T_2 say the same about Σ'

Let Σ be a signature, T_1, T_2 theories.

- T_1 and T_2 are Σ-model inseparable if

$$\{M_{|\Sigma} \mid M \models T_1\} \quad = \quad \{M_{|\Sigma} \mid M \models T_2\}$$

- Let $\mathcal{QL}$ be a query language of interest. T_1 and T_2 are Σ-inseparable w.r.t. $\mathcal{QL}$ if

$$T_1 \models \varphi \quad \Leftrightarrow \quad T_2 \models \varphi$$

  for all $\varphi \in \mathcal{QL}$ with $\mathrm{sig}(\varphi) \subseteq \Sigma$.

If $T_1 \subseteq T_2$ and $\Sigma = \mathrm{sig}(T_1)$, then

$$\text{inseparability} = \text{conservative extension.}$$

Description Logic: $\mathcal{ALC}$

Concepts are defined as

$$C, D := \quad A \mid C \sqcap D \mid \neg C \mid \exists r.C \mid \forall r.C.$$

- Description Logic:

$$\text{Human} \sqcap \neg\text{Female} \sqcap \exists\text{child}.\top \sqcap \forall\text{child}.\text{Male}$$

- Modal Logic:

$$\text{Human} \wedge \neg\text{Female} \wedge \Diamond_{\text{child}}\top \wedge \Box_{\text{child}}\text{Male}$$

- First-order Logic:

$$\text{Human}(x) \wedge \neg\text{Female}(x) \wedge \exists y.\text{child}(x, y) \wedge \forall y.\text{child}(x, y) \rightarrow \text{Male}(y)$$

A sentence is an implication $C_1 \sqsubseteq C_2$ between concepts.

Definition

$M \models C_1 \sqsubseteq C_2$ iff $M \models \forall x.C_1(x) \rightarrow C_2(x)$.

Ontologies in Description Logic

An $\mathcal{ALC}$-TBox (*ontology*) is a finite set of sentences $C_1 \sqsubseteq C_2$.

$$
\begin{array}{rcl}
\text{Cystic\_Fibrosis} & \sqsubseteq & \text{Fibrosis} \sqcap \exists \text{located\_In.Pancreas} \sqcap \exists \text{has\_Origin.Genetic\_Origin} \\
\text{Genetic\_Fibrosis} & \equiv & \text{Fibrosis} \sqcap \exists \text{has\_Origin.Genetic\_Origin} \\
\text{Genetic\_Fibrosis} & \sqsupseteq & \text{Fibrosis} \sqcap \exists \text{located\_In.Pancreas} \\
\text{Genetic\_Fibrosis} & \sqsubseteq & \text{Genetic\_Disorder} \\
\text{DEFBI\_Gene} & \sqsubseteq & \text{Immuno\_Protein\_Gene} \sqcap \exists \text{associated\_With.Cystic\_Fibrosis}
\end{array}
$$

Theorem

Deciding whether $T \models C \sqsubseteq D$ is

- *ExpTime-complete for $\mathcal{ALC}$;*
- *PTime-complete for $\mathcal{EL}$ (only $\sqcap$ and $\exists$).*

Inseparability and conservative extensions

Let Σ be a signature. We are interested in

- versioning: check whether T_1 and T_2 are Σ-inseparable;
- module extraction: compute minimal $M \subseteq T$ such that M and T are Σ-inseparable.
- uniform interpolation: compute finite T_Σ such that T_Σ uses Σ only and T and T_Σ are Σ-inseparable.

Problem: decide Σ-inseparability for $\mathcal{ALC}$ and $\mathcal{EL}$.

Model inseparability/conservative extension

Theorem

In $\mathcal{EL}$ and $\mathcal{ALC}$, deciding model inseparability/conservative extension is as hard as monadic second-order logic.

Proof. ($\mathcal{ALC}$) It is sufficient to show this for validity of

$$\exists \vec{p} \varphi \rightarrow \exists \vec{p} \psi$$

for modal logic formulas φ and ψ.

(Thomason, 1975) Validity of $\forall \vec{p}\, \varphi \rightarrow \forall \vec{p}\, \psi$ is as hard as monadic second-order logic.

Model inseparability/conservative extensions: unary predicates

Theorem

Assume Σ consists of unary predicates only. Then Σ-model inseparability is coNExpTime^{NP}-complete, in $\mathcal{EL}$ and $\mathcal{ALC}$.

Upper bound: Guess a counterexample M of exponential size and call an NP-oracle to check whether it is a counterexample.
Lower bound: Reduction of complement of succinct version of Cert3Col: given an undirected graph with edges labelled by disjunction of two literals, check whether there is a truth assignment such that the resulting graph is not 3-colorable.

Concept vs. model conservative extensions in $\mathcal{ALC}$

Definition

Let $T_1 \subseteq T_2$ be $\mathcal{ALC}$-TBoxes. T_2 is a concept conservative extension of T_1 iff

$$T_2 \models C \sqsubseteq D \quad \Leftrightarrow \quad T_1 \models C \sqsubseteq D,$$

whenever $\mathrm{sig}(C \sqsubseteq D) \subseteq \mathrm{sig}(T_1)$.

- $T_1 = \{\top \sqsubseteq \exists r.\top\}$;
- $T_2 = T_1 \cup \{\top \sqsubseteq \exists r.A \sqcap \exists r.\neg A\}$;
- T_2 is not a model conservative extension of T_1;
- T_2 is a concept conservative extension of T_1 in $\mathcal{ALC}$.

Characterization of concept conservative extensions in $\mathcal{ALC}$

Assume a characterization $\mathcal{ALC}$-equivalence using bisimulations. Let $T_2 \supseteq T_1$ be $\mathcal{ALC}$-TBoxes. To see whether T_2 is a conservative extension of T_1 do for

- model conservative extension: check validity of

$$(\bigwedge T_1) \rightarrow \exists \mathsf{new}(T_2)(\bigwedge T_2)$$

  (second-order quantifier).
- for concept conservative extension: check validity of

$$(\bigwedge T_1) \rightarrow \exists^{bisim} \mathsf{new}(T_2) \bigwedge (T_2)$$

  (bisimulation quantifier).

Characterizing logical equivalence in $\mathcal{ALC}$: Bisimulation

Given a signature Σ and two models M_1 and M_2, a relation $\rho \subseteq \Delta_1 \times \Delta_2$ is a Σ-bisimulation iff

- $(v_1, v_2) \in \rho$ implies $v_1 \in A^{M_1}$ iff $v_2 \in A^{M_2}$, for $A \in \Sigma$;
- If $(v_1, v_2) \in \rho$ and $(v_1, v_1') \in r^{M_1}$, then there exists v_2' with $(v_2, v_2') \in r^{M_2}$ and $(v_1', v_2') \in \rho$, for $r \in \Sigma$.
- vice versa.

$(M_1, w_1) \sim_\Sigma (M_2, w_2)$ (w_1 and w_2 are Σ-bisimilar) if there is a Σ-bisimulation ρ with $(w_1, w_2) \in \rho$.

Logical equivalence in $\mathcal{ALC}$

Theorem

For finite models the following are equivalent:

- $(M_1, w_1) \sim_\Sigma (M_2, w_2)$;
- w_1 *and* w_2 *satisfy the same* Σ*-concepts; i.e., for all* C *over* Σ*:*

$$w_1 \in C^{M_1} \Leftrightarrow w_2 \in C^{M_2}.$$

(Does not hold for all infinite models.)

Deciding concept conservative extensions in $\mathcal{ALC}$

Theorem

Concept conservative extensions in $\mathcal{ALC}$ is 2ExpTime-complete.

Proof (Upper bound, using automata)

- Check satisfiability of $T_1 \wedge \neg\exists^{bisim}\text{new}(T_2)T_2$.
- Construct μ*-automaton* (Janin, Walukiewicz/Wilke) accepting exactly the models of $T_1 \wedge \neg\exists^{bisim}\text{new}(T_2)T_2$. Then check emptiness.

Concept conservative extensions in $\mathcal{EL}$

Theorem

Concept conservative extensions in $\mathcal{EL}$ is ExpTime-complete. (Characterization of $\mathcal{EL}$-logical equivalence using simulations instead of bisimulations.)

Uniform Interpolation

Let T be a $\mathcal{EL}/\mathcal{ALC}$-TBox, Σ a signature. A $\mathcal{EL}/\mathcal{ALC}$-TBox T_Σ is called a uniform interpolant of T w.r.t. Σ if the following holds:

- $\text{sig}(T_\Sigma) \subseteq \Sigma$;
- T and T_Σ are Σ-inseparable w.r.t. $\mathcal{EL}/\mathcal{ALC}$.

Uniform interpolants do not always exist

Let

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

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

- if $d_0 \in A_0^M$, then exists a sequence $d_0 r^M d_1 r^M d_2 r^M \dots r^M d_n$ (for all $n > 0$)

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

Deciding the existence of uniform interpolants

Theorem

In $\mathcal{ALC}$, the problem of deciding the existence of uniform interpolants is 2ExpTime-complete.

For $\mathcal{EL}$ decidability is open.

Support from logic for developing and using ontologies

Let Σ be a signature (a subject matter). We are interested in

- versioning: check whether T_1 and T_2 are Σ-inseparable;
- module extraction: compute minimal $M \subseteq T$ such that M and T are Σ-inseparable;
- uniform interpolation: compute finite T_Σ such that T_Σ uses Σ only and T and T_Σ are Σ-inseparable.

Problem: Computationally hard even for $\mathcal{EL}$-TBoxes:

- Model conservative extensions undecidable;
- $\mathcal{EL}$-conservative extensions is ExpTime-complete.
- No procedures known for computing uniform interpolants.

We consider $\mathcal{EL}$-TBoxes of a particular form.

$\mathcal{EL}$-terminologies

Definition

An $\mathcal{EL}$-TBox T is a $\mathcal{EL}$-terminology if axioms are of the form

- $A \equiv C$ or $A \sqsubseteq C$,

where A is a concept name and no A occurs more than once on the left hand side.
A $\mathcal{EL}$-terminology T is acyclic if no concept name refers to itself along definitions.

SNOMED CT is an acyclic $\mathcal{EL}$-terminology (with some additional constructors).

Model conservative extensions

Theorem

For acyclic $\mathcal{EL}$-terminologies model conservative extensions are decidable in polynomial time.

Theorem

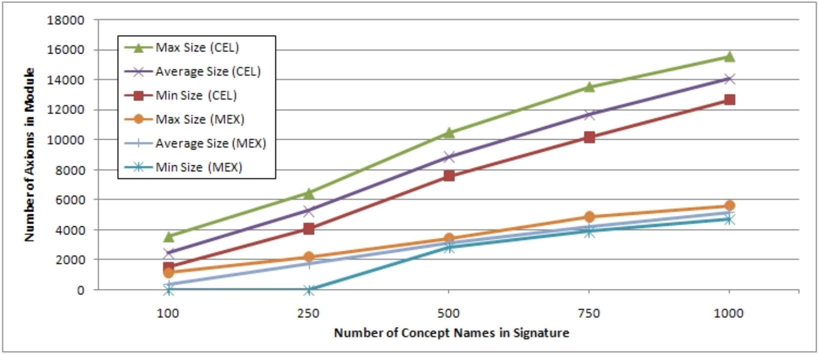
Let T be an acyclic $\mathcal{EL}$-terminology and Σ a signature. Then one can compute the minimal subset M of T such that

- $\Sigma \subseteq \mathsf{sig}(M)$ *and*
- T *is a model-conservative extension of* M.

in polynomial time.

Experiment: Extraction of modules from SNOMED CT

- We use a prototype implementation MEX.
- Σ — randomly selected from SNOMED CT.
- 1000 samples for each signature size
- with role box



Uniform interpolation: 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.

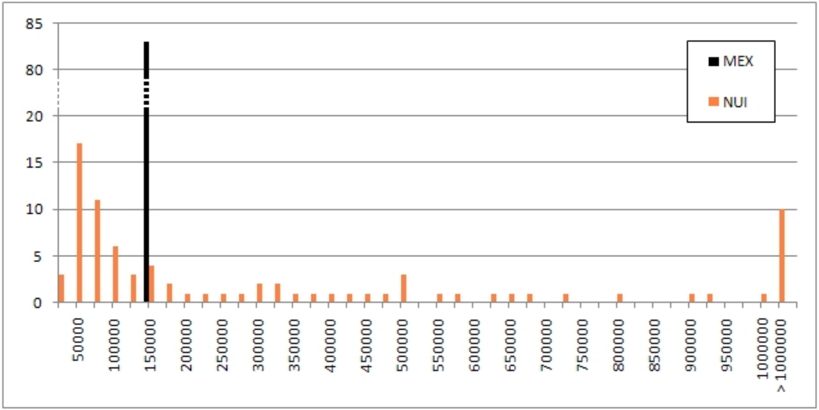
Computing uniform interpolants for SNOMED CT and NCI: success rate

- We use implementation NUI
- 100 randomly generated signatures.

$\lvert\Sigma\rvert$	SNOMED CT	$\lvert\Sigma\rvert$	NCI
2 000	100.0%	5 000	97.0%
3 000	92.2%	10 000	81.1%
4 000	67.0%	15 000	72.0%
5 000	60.0%	20 000	59.2%

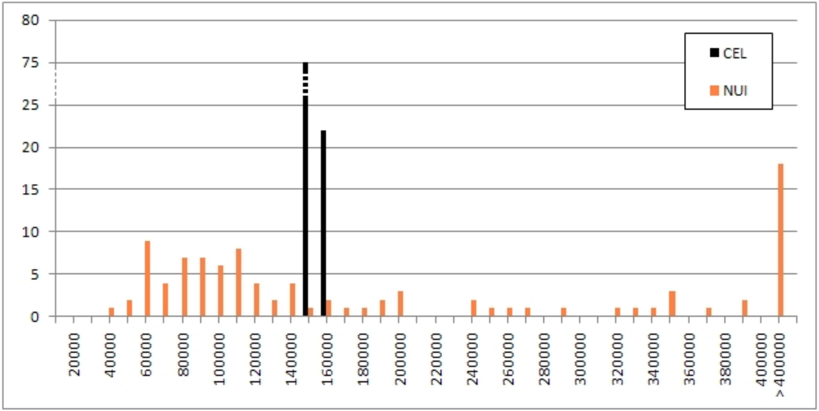
Comparing the size of MEX-modules and Σ-interpolants

- Size distribution of MEX-modules and Σ-interpolants of SNOMED CT wrt. signatures containing 3 000 concept names and 20 role names



Comparing the size of modules and Σ-interpolants

- Size distribution of CEL-modules and Σ-interpolants of NCI wrt. signatures containing 7 000 concept names and 20 role names



Conclusion

Apply other notions from logic:

- Interpretations between theories vs mappings.
- Abstract model theory for ontology languages.