The description logic $\mathcal{EL}$: the terminological part

DL architecture

Knowledge Base (KB)

TBox (terminological box, schema)

$$\text{Man} \equiv \text{Human} \sqcap \text{Male}$$
$$\text{HappyFather} \equiv \text{Man} \sqcap \exists\text{hasChild}$$
...

ABox (assertion box, data)

$$\text{john} : \text{Man}$$
$$(\text{john}, \text{mary}) : \text{hasChild}$$
...

Inference System

Interface

$\mathcal{EL}$ (syntax)

- **Language for $\mathcal{EL}$ concepts (classes):**
  - concept names $A_0, A_1, \ldots$ (e.g., Person, Female, ...)
  - role names $r_0, r_1, \ldots$ (e.g., hasChild, loves, ...)
  - the concept $\top$ (often called "thing")
  - the concept constructor $\sqcap$ (often called intersection, conjunction, or simply "and").
  - the concept constructor $\exists$ (often called existential restriction).

- **$\mathcal{EL}$ concepts are defined inductively:**
  - all concept names are $\mathcal{EL}$ concepts
  - $\top$ is a $\mathcal{EL}$ concept
  - if C and D are $\mathcal{EL}$ concepts and r is a role name, then

    $$(C \sqcap D), \quad \exists r.C$$

    are $\mathcal{EL}$ concepts.

Examples

- $\mathbf{Person} \sqcap \mathbf{Female}$ (a woman),
- $\mathbf{Person} \sqcap \exists \mathbf{hasChild.Person}$ (a person who has a child),
- $\mathbf{Person} \sqcap \exists \mathbf{hasChild.Person} \sqcap \exists \mathbf{hasParent.Person}$ (a person who has a child and has a parent),
- $\mathbf{Person} \sqcap \exists \mathbf{hasChild.}(\mathbf{Person} \sqcap \mathbf{Female})$ (a person who has a child who is a woman),
- $\mathbf{Person} \sqcap \exists \mathbf{hasChild.Person} \sqcap \mathbf{Female}$ (a woman who has a child),
- $\mathbf{Person} \sqcap \exists \mathbf{hasChild.}\top$ (a person who has a child),
- $\mathbf{Person} \sqcap \exists \mathbf{hasChild.} \exists \mathbf{hasChild.}\top$ (a person who has a grandchild).

Concept definitions in $\mathcal{EL}$

Let A be a concept name and C a $\mathcal{EL}$ concept. Then

- $A \equiv C$ is a $\mathcal{EL}$ concept definition. C describes necessary and sufficient conditions for being an A. We sometimes read this as "A is equivalent to C".

- $A \sqsubseteq C$ is a primitive $\mathcal{EL}$ concept definition. C describes necessary conditions for being an A. We sometimes read this as "A is subsumed by C".

Examples:

- $\mathbf{Father} = \mathbf{Person} \sqcap \mathbf{Male} \sqcap \exists \mathbf{hasChild}.\top$.

- $\mathbf{Student} = \mathbf{Person} \sqcap \exists \mathbf{is\_registered\_at}.\mathbf{University}$.

- $\mathbf{Father} \sqsubseteq \mathbf{Person}$.

- $\mathbf{Father} \sqsubseteq \exists \mathbf{hasChild}.\top$.

$\mathcal{EL}$ terminology

A **$\mathcal{EL}$ terminology $\mathcal{T}$** is a finite set of definitions of the form

$$A \equiv C, \quad A \sqsubseteq C$$

such that no concept name occurs more than once on the left hand side of a definition. So, in a terminology it is impossible to have two distinct definitions:

- $\textbf{University} \equiv \textbf{Institution} \sqcap \exists \textbf{grants.academicdegree}$
- $\textbf{University} \equiv \textbf{Institution} \sqcap \exists \textbf{supplies.higher\_education}$

However, we can have cyclic definitions such as

$$\textbf{Human\_being} \equiv \exists \textbf{has\_parent.Human\_being}$$

A **acyclic $\mathcal{EL}$ terminology $\mathcal{T}$** is a $\mathcal{EL}$ terminology that does not not contain (not even indirect) cyclic definitions.

Example: SNOMED CT

- Comprehensive healthcare terminology with approximately 400 000 definitions (400 000 concept names and 60 role names)

- Almost (except inclusions between role names) an acyclic $\mathcal{EL}$ terminology

- 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.

- The widespread adoption of SNOMED CT across all NHS clinical systems is a strategic goal of the NHS Connecting for Health IT programme.

SNOMED CT Snippet

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

$\mathcal{EL}$ concept inclusion (CI)

Let C and D be $\mathcal{EL}$ concepts. Then

- $C \sqsubseteq D$ is called a **$\mathcal{EL}$ concept inclusion**. It states that every C **is-a** D. We also say that C is subsumed by D or that D subsumes C. Sometimes we also say that C is included in D.
- $C \equiv D$ is an abbreviation for the two concept inclusions $C \sqsubseteq D$ and $D \sqsubseteq C$. We sometimes read this as "C and D are equivalent".

Examples:

- $\textbf{Disease} \sqcap \exists \textbf{has\_location}.\textbf{Heart} \sqsubseteq \textbf{NeedsTreatment}$
- $\exists \textbf{student\_of}.\textbf{ComputerScience} \sqsubseteq \textbf{Human\_being} \sqcap \exists \textbf{knows}.\textbf{Programming\_Language}$

$\mathcal{EL}$ TBox

A $\mathcal{EL}$ **TBox** is a finite set $\mathcal{T}$ of $\mathcal{EL}$ concept inclusions $C \sqsubseteq D$ (we use $C \equiv D$ as an abbreviation). Note the following inclusions:

$$\text{acyclic terminology} \quad \Rightarrow \quad \text{terminology} \quad \Rightarrow \quad \text{TBox}$$

Example:

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

How are TBoxes (eg, SNOMED CT) used?

The **concept hierarchy** induced by a TBox $\mathcal{T}$ is defined as

$$\{A \sqsubseteq B \mid A, B \text{ concept names in } \mathcal{T} \text{ and } \mathcal{T} \text{ implies } A \sqsubseteq B\}$$

Eg, the concept hierarchy induced by the SNOMED CT snippet above is

$$\begin{gathered}\mathsf{EntireDistalEpiphysisOfFemur}\\ \sqsubseteq \\ \mathsf{StructureOfDistalEpiphysisOfFemur}\\ \sqsubseteq \\ \mathsf{DistalFemurPart}\\ \sqsubseteq \\ \mathsf{BoneStructureOfDistalFemur}\\ \sqsubseteq \\ \mathsf{FemurPart}\end{gathered}$$

Standard application of SNOMED CT based on concept hierarchy

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

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

Problem: we do not yet have a precise definition of what it means that $A \sqsubseteq B$ follows from $\mathcal{T}$ (or is implied by $\mathcal{T}$). So: we do not have a precise definition of the concept hierarchy induced by a TBox

$\mathcal{EL}$ (semantics)

- An **interpretation** is a structure $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ in which
  - $\Delta^{\mathcal{I}}$ is the **domain** (a non-empty set)
  - $\cdot^{\mathcal{I}}$ is an **interpretation function** that maps:
    - every concept name A to a subseteq $A^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ $\quad (A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}})$
    - every role name r to a binary relation $r^{\mathcal{I}}$ over $\Delta^{\mathcal{I}}$ $\quad (r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}})$
- The interpretation $C^{\mathcal{I}}$ of an arbitrary $\mathcal{EL}$ concept C in $\mathcal{I}$ is defined inductively:
  - $(\top)^{\mathcal{I}} = \Delta^{\mathcal{I}}$
  - $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$
  - $(\exists r.C)^{\mathcal{I}} = \{x \in \Delta^{\mathcal{I}} \mid \text{ exists } y \in \Delta^{\mathcal{I}} \text{ such that } (x, y) \in r^{\mathcal{I}} \text{ and } y \in C^{\mathcal{I}}\}$

Example

Let $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where

- $\Delta^{\mathcal{I}} = \{a, b, c, d, e, f\}$;
- $\mathbf{Person}^{\mathcal{I}} = \{a, b, c, d, f\}$; $\mathbf{Female}^{\mathcal{I}} = \{a, b, c, e\}$;
- $\mathbf{hasChild}^{\mathcal{I}} = \{(a, b), (b, c), (d, e), (f, f)\}$.

Compute:

- $(\mathbf{Person} \sqcap \mathbf{Female})^{\mathcal{I}}$,
- $(\mathbf{Person} \sqcap \exists \mathbf{hasChild}.\mathbf{Person})^{\mathcal{I}}$,
- $(\mathbf{Person} \sqcap \exists \mathbf{hasChild}.(\mathbf{Person} \sqcap \mathbf{Female})^{\mathcal{I}}$,
- $(\mathbf{Person} \sqcap \exists \mathbf{hasChild}.\mathbf{Person} \sqcap \mathbf{Female})^{\mathcal{I}}$,
- $(\mathbf{Person} \sqcap \exists \mathbf{hasChild}.\top)^{\mathcal{I}}$,
- $(\mathbf{Person} \sqcap \exists \mathbf{hasChild}.\exists \mathbf{hasChild}.\top)^{\mathcal{I}}$.

Semantics: when is a concept inclusion true in an interpretation?

Let $\mathcal{I}$ be an interpretation, $C \sqsubseteq D$ a concept inclusion, and $\mathcal{T}$ a TBox.

- We set $\mathcal{I} \models C \sqsubseteq D$ if, and only if, $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. In words:
  - $\mathcal{I}$ satisfies $C \sqsubseteq D$ or
  - $C \sqsubseteq D$ is true in $\mathcal{I}$ or
  - $\mathcal{I}$ is a model of $C \sqsubseteq D$.
- We set $\mathcal{I} \models C \equiv D$ if, and only if, $C^{\mathcal{I}} = D^{\mathcal{I}}$
- We set $\mathcal{I} \models \mathcal{T}$ if, and only if, $\mathcal{I} \models E \sqsubseteq F$ for all $E \sqsubseteq F$ in $\mathcal{T}$. In words:
  - $\mathcal{I}$ satisfies $\mathcal{T}$ or
  - $\mathcal{I}$ is a model of $\mathcal{T}$.

Semantics: when does a concept inclusion follow from a TBox?

Let $\mathcal{T}$ be a TBox and $C \sqsubseteq D$ a concept inclusion. We say that **$C \sqsubseteq D$ follows from $\mathcal{T}$** if, and only if, every model of $\mathcal{T}$ is a model of $C \sqsubseteq D$.

Instead of saying that $C \sqsubseteq D$ follows from $\mathcal{T}$ we often write

- $\mathcal{T} \models C \sqsubseteq D$ or
- $C \sqsubseteq_{\mathcal{T}} D$.

Example: let MED be the $\mathcal{EL}$ TBox

$$
\begin{array}{rcl}
\textbf{Pericardium} & \sqsubseteq & \textbf{Tissue} \sqcap \exists \textbf{cont\_in}.\textbf{Heart} \\
\textbf{Pericarditis} & \sqsubseteq & \textbf{Inflammation} \sqcap \exists \textbf{has\_loc}.\textbf{Pericardium} \\
\textbf{Inflammation} & \sqsubseteq & \textbf{Disease} \sqcap \exists \textbf{acts\_on}.\textbf{Tissue} \\
\textbf{Disease} \sqcap \exists \textbf{has\_loc}.\exists \textbf{cont\_in}.\textbf{Heart} & \sqsubseteq & \textbf{Heartdisease} \sqcap \textbf{NeedsTreatment}
\end{array}
$$

Pericarditis needs treatment if, and only if, $\textbf{Percarditis} \sqsubseteq_{\textsf{MED}} \textbf{NeedsTreatment}$.

Deciding whether $C \sqsubseteq_{\mathcal{T}} D$ for $\mathcal{EL}$-TBoxes $\mathcal{T}$.

We give a polynomial time (tractable) algorithm deciding whether $C \sqsubseteq_{\mathcal{T}} D$

The algorithm actually decides whether $A \sqsubseteq_{\mathcal{T}} B$ for concept names A and B in $\mathcal{T}$. This is sufficient because the following two conditions are equivalent:

- $C \sqsubseteq_{\mathcal{T}} D$
- $A \sqsubseteq_{\mathcal{T}'} B$ for fresh concept names A and B and the TBox

$$\mathcal{T}' = \mathcal{T} \cup \{A \equiv C, B \equiv D\}$$

Pre-processing

A $\mathcal{EL}$-TBox is in *normal form* if it consists of inclusions of the form

(sform) $A \sqsubseteq B$, where A and B are concept names;

(cform) $A_1 \sqcap A_2 \sqsubseteq B$, where A_1, A_2, B are concept names;

(rform) $A \sqsubseteq \exists r.B$, where A, B are concept names;

(lform) $\exists r.A \sqsubseteq B$, where A, B are concept names.

Given a $\mathcal{EL}$-Box $\mathcal{T}$, one can compute in polynomial time a TBox $\mathcal{T}'$ in normal form such that for all concept names A, B in $\mathcal{T}$:

$$A \sqsubseteq_{\mathcal{T}} B \quad \Leftrightarrow \quad A \sqsubseteq_{\mathcal{T}'} B.$$

Algorithm for Pre-processing (main steps)

Given a TBox $\mathcal{T}$, apply the following rules exhaustively:

- Replace each $C_1 \equiv C_2$ by $C_1 \sqsubseteq C_2$ and $C_2 \sqsubseteq C_1$;
- Replace each $C \sqsubseteq C_1 \sqcap C_2$ by $C \sqsubseteq C_1$ and $C \sqsubseteq C_2$;
- If $\exists r.C$ occurs in $\mathcal{T}$ and C is complex, replace C everywhere by a fresh concept name X_C and add $X_C \sqsubseteq C$ and $C \sqsubseteq X_C$ to the TBox.
- If $A_1 \sqcap \cdots \sqcap A_n \sqcap \exists r_1.B_1 \sqcap \cdots \sqcap \exists r_m.B_m \sqsubseteq C$ in $\mathcal{T}$, remove it, take a new concept name X and add

$$A_1 \sqcap \cdots \sqcap A_n \sqcap X \sqsubseteq C \quad \exists r_1.B_1 \sqcap \cdots \sqcap \exists r_m.B_m \sqsubseteq X$$

- If $\exists r_1.B_1 \sqcap \cdots \sqcap \exists r_m.B_m \sqsubseteq \exists r.B$ in $\mathcal{T}$, remove it, take a new concept name X and add

$$\exists r_1.B_1 \sqcap \cdots \sqcap \exists r_m.B_m \sqsubseteq X \quad X \sqsubseteq \exists r.B$$

Pre-Processing: Example

Consider $\mathcal{T}$:

$$\begin{array}{rcl} A_0 & \sqsubseteq & B \sqcap \exists r.B' \\ A_1 \sqcap \exists r.B & \sqsubseteq & A_2 \end{array}$$

Step 1 gives:

$$\begin{array}{rcl} A_0 & \sqsubseteq & B \\ A_0 & \sqsubseteq & \exists r.B' \\ A_1 \sqcap \exists r.B & \sqsubseteq & A_2 \end{array}$$

Step 2 gives:

$$\begin{array}{rcl} A_0 & \sqsubseteq & B \\ A_0 & \sqsubseteq & \exists r.B' \\ A_1 \sqcap X & \sqsubseteq & A_2 \\ \exists r.B & \sqsubseteq & X \end{array}$$

Pre-Processing applied to Example MED

$$
\begin{array}{rcl}
\textsf{Pericardium} & \sqsubseteq & \textsf{Tissue} \\
\textsf{Pericardium} & \sqsubseteq & Y \\
\textsf{Pericarditis} & \sqsubseteq & \textsf{Inflammation} \\
\textsf{Pericarditis} & \sqsubseteq & \exists \textsf{has\_loc.Pericardium} \\
\textsf{Inflammation} & \sqsubseteq & \textsf{Disease} \\
\textsf{Inflammation} & \sqsubseteq & \exists \textsf{acts\_on.Tissue} \\
\textsf{Disease} \sqcap X & \sqsubseteq & \textsf{Heartdisease} \\
\textsf{Disease} \sqcap X & \sqsubseteq & \textsf{NeedsTreatment}
\end{array}
$$

$$
\exists \textsf{has\_loc.}Y \sqsubseteq X \quad \exists \textsf{cont\_in.Heart} \sqsubseteq Y \quad Y \sqsubseteq \exists \textsf{cont\_in.Heart}
$$

Algorithm deciding $A \sqsubseteq_{\mathcal{T}} B$: Intuition

Given $\mathcal{T}$ in normal form, we compute functions:

- S maps every concept name A from $\mathcal{T}$ a set of concept names B;
- R maps every role name r from $\mathcal{T}$ to a set of pairs (B_1, B_2) of concept names.

We will have $A \sqsubseteq_{\mathcal{T}} B$ if, and only if, $B \in S(A)$. Intuitively, we construct an interpretation $\mathcal{I}$ with

- $\Delta^{\mathcal{I}}$ is the set of concept names in $\mathcal{T}$.
- $A^{\mathcal{I}}$ is the set of all B such that $A \in S(B)$;
- $r^{\mathcal{I}}$ is the set of all $(A, B) \in R(r)$.

This will be a model of $\mathcal{T}$ and $A \sqsubseteq_{\mathcal{T}} B$ if, and only if, $A \in B^{\mathcal{I}}$.

Algorithm

Input: $\mathcal{T}$ in normal form. Initialise: $S(A) = \{A\}$ and $R(r) = \emptyset$ for A and r in $\mathcal{T}$. Apply the following four rules to S and R exhaustively:

(simpleR) If $A' \in S(A)$ and $A' \sqsubseteq B \in \mathcal{T}$ and $B \notin S(A)$, then

$$S(A) := S(A) \cup \{B\}.$$

(conjR) If $A_1, A_2 \in S(A)$ and $A_1 \sqcap A_2 \sqsubseteq B \in \mathcal{T}$ and $B \notin S(A)$, then

$$S(A) := S(A) \cup \{B\}.$$

(rightR) If $A' \in S(A)$ and $A' \sqsubseteq \exists r.B \in \mathcal{T}$ and $(A, B) \notin R(r)$, then

$$R(r) := R(r) \cup \{(A, B)\}.$$

(leftR) If $(A, B) \in R(r)$ and $B' \in S(B)$ and $\exists r.B' \sqsubseteq A' \in \mathcal{T}$ and $A' \notin S(A)$, then

$$S(A) := S(A) \cup \{A'\}.$$

Example

$$\begin{aligned} A_0 &\sqsubseteq \exists r.B \\ B &\sqsubseteq E \\ \exists r.E &\sqsubseteq A_1 \end{aligned}$$

Initialise: $S(A_0) = \{A_0\}$, $S(A_1) = \{A_1\}$, $S(B) = \{B\}$, $S(E) = \{E\}$, $R(r) = \emptyset$.

- Application of (rightR) and axiom 1 gives: $R(r) = \{(A_0, B)\}$;
- Application of (simpleR) and axiom 2 gives: $S(B) = \{B, E\}$;
- Application of (leftR) and axiom 3 gives: $S(A_0) = \{A_0, A_1\}$;
- No more rules are applicable.

Thus, $R(r) = \{(A_0, B)\}$, $S(B) = \{B, E\}$, $S(A_0) = \{A_0, A_1\}$ and no changes for the remaining values. We obtain $A_0 \sqsubseteq_{\mathcal{T}} A_1$.

Fragment of MED

$$
\begin{aligned}
\textbf{Pericardium (Pm)} &\sqsubseteq Y \\
\textbf{Pericarditis (Ps)} &\sqsubseteq \textbf{Inflammation (Inf)} \\
\textbf{Ps} &\sqsubseteq \exists\textbf{has\_loc.Pm} \\
\textbf{Inf} &\sqsubseteq \textbf{Disease (Dis)} \\
\textbf{Disease} \sqcap X &\sqsubseteq \textbf{NeedsTreatment} \\
\exists\textbf{has\_loc.}Y &\sqsubseteq X
\end{aligned}
$$

Partial run of the algorithm (showing that $\textbf{Ps} \sqsubseteq_{\textsf{MED}} \textbf{NeedsTreatment}$):

- Applications of (simpleR) give: $S(\textbf{Pm}) = \{Y, \textbf{Pm}\}$, $S(\textbf{Ps}) = \{\textbf{Inf}, \textbf{Ps}, \textbf{Dis}\}$;
- Application of (rightR) give: $R(\textbf{has\_loc}) = \{(\textbf{Ps}, \textbf{Pm})\}$,
- Application of (leftR) gives: $S(\textbf{Ps}) = \{\textbf{Inf}, \textbf{Ps}, \textbf{Dis}, X\}$
- Application of (conjR) gives: $S(\textbf{Ps}) = \{\textbf{Inf}, \textbf{Ps}, \textbf{Dis}, X, \textbf{NeedsTreatment}\}$

Analysing the output of the algorithm

Let $\mathcal{T}$ be in normal form and S, R the output of the algorithm.

Theorem. For all concept names A, B in $\mathcal{T}$: $A \sqsubseteq_{\mathcal{T}} B$ if, and only if, $B \in S(A)$.

In fact, the following holds: Define an interpretation $\mathcal{I}$ by

- $\Delta^{\mathcal{I}}$ is the set of concept names in $\mathcal{T}$.
- $A^{\mathcal{I}}$ is the set of all B such that $A \in S(B)$;
- $r^{\mathcal{I}}$ is the set of all $(A, B) \in R(r)$.

Then

- $\mathcal{I}$ satisfies $\mathcal{T}$ and
- for all concept names A from $\mathcal{T}$ and $\mathcal{EL}$-concepts C:

$$A \sqsubseteq_{\mathcal{T}} C \quad \Leftrightarrow \quad A \in C^{\mathcal{I}}.$$