CEX and MEX: Logical Diff and Logic-based Module Extraction in a Fragment of OWL

B. Konev, C.Lutz, D.Walther, and F.Wolter

Liverpool University, UK
TU Dresden, Germany

April 1, 2008
Motivation: SNOMED CT

- **SNOMED CT** — the Systematised Nomenclature of Medicine (Clinical Terms).
- ~ 400,000 terms
- Simple structure:

  \[
  \text{Hindbrain\_hernia\_headache} \sqsubseteq \\
  \text{Headache\_disorder} \sqcap \exists \text{Due\_to\_Cerebral\_herniation}
  \]

- $\mathcal{E}L$ representation

NB: $\mathcal{E}L$ will be a part of the updated OWL (??)
Understanding Changes in Terminologies

Suppose the following

\[
\begin{align*}
\text{Neck\_injection} & \sqsubseteq \text{Operation} \\
\text{Neck\_operation} & \sqsubseteq \text{Operation} \\
\text{Removal\_f.b\_.from\_neck} & \equiv \text{Neck\_operation} \sqcap \text{Removal\_foreign\_body}
\end{align*}
\]

is refined as

\[
\begin{align*}
\text{Neck\_injection} & \sqsubseteq \text{Neck\_operation} \\
\text{Neck\_operation} & \sqsubseteq \text{Operation} \\
\text{Removal\_f.b\_.from\_neck} & \equiv \text{Neck\_operation} \sqcap \text{Removal\_foreign\_body}
\end{align*}
\]

The refined terminology implies

- \( \text{Neck\_injection} \sqsubseteq \text{Neck\_operation} \)
- \( \text{Neck\_injection} \sqcap \text{Removal\_foreign\_body} \sqsubseteq \text{Removal\_f.b\_.from\_neck} \)
What’s the difference between different versions of a terminology?

- What’s the difference over some signature $\Sigma$?

- $T_1, T_2 - \mathcal{EL}$ terminologies
- $\Sigma -$ signature

$$\text{diff}_\Sigma(T_1, T_2) = \left\{ C \sqsubseteq D \mid \begin{array}{l} T_1 \not\models C \sqsubseteq D \\ T_2 \models C \sqsubseteq D \\ \text{sig}(C \sqsubseteq D) \subseteq \Sigma \end{array} \right\}$$

- We give a compact representation of $\text{diff}_\Sigma(T_1, T_2)$
- Polytime algorithm
diff(SNOMED CT’05, SNOMED CT’06) on
\[ \Sigma = \text{sig}(SNOMED\ CT'05) \cap \text{sig}(SNOMED\ CT'06) \]
- 689 seconds
- \(|\text{diffL}_\Sigma| + |\text{diffR}_\Sigma| = 162010\)
- Class hierarchy comparison misses **32475** of them!
SNOMED CT’05 vs SNOMED CT’06: Fragments

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

| Size of $\Sigma$ | CEX: diff(SNOMED CT’05,SNOMED CT’06) | Time (Sec.) | Memory (MByte) | $|\text{diffL}_\Sigma|$ | $|\text{diffR}_\Sigma|$ |
|-----------------|---------------------------------------|--------------|----------------|----------------|----------------|
| 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        |

- Role box ignored
Semantic Modules

- ~ 10–15 minutes is fine, but we want better
- Time depends more on the terminology size than on \( \Sigma \)
- Extract modules!
  - Module should be a “replacement” for the terminology

**Theorem**

- \( T \) — acyclic EL terminology
- *No trivial axioms* \( A \equiv \top \), etc in \( T \)

The smallest semantic module can be computed in polytime
## CEX on MEX

<table>
<thead>
<tr>
<th>Size of Σ</th>
<th>CEX: diff(SPOMED CT’05, SPOMED CT’06)</th>
<th>CEX: diff(Mod’05, Mod’06)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time (Sec.)</td>
<td>Memory (MByte)</td>
</tr>
<tr>
<td>100</td>
<td>513.1</td>
<td>1 393.7</td>
</tr>
<tr>
<td>1 000</td>
<td>512.4</td>
<td>1 394.6</td>
</tr>
<tr>
<td>10 000</td>
<td>517.7</td>
<td>1 424.3</td>
</tr>
<tr>
<td>100 000</td>
<td>559.8</td>
<td>1 473.2</td>
</tr>
</tbody>
</table>

8 / 13
Consider

\[
\begin{align*}
\text{Hindbrain\_hernia\_headache} & \sqsubseteq \\
\text{Headache\_disorder} & \sqcap \exists \text{Due\_to\_Cerebral\_herniation} \\
\text{Removal\_f.b.\_from\_neck} & \equiv \text{Neck\_operation} \sqcap \text{Removal\_foreign\_body}
\end{align*}
\]

- \( \Sigma = \{ \text{Hindbrain\_hernia\_headache, Removal\_f.b.\_from\_neck} \} \)
- \( M = \emptyset \) (nothing!)

Many module extraction techniques would select both axioms!

- definition closed
Definition-Closed vs Semantic Modules: Size

- $\Sigma$ — randomly selected from SNOMED CT’05
- 1000 samples for each signature size
- with role box (under simplifying assumptions)
Definition-Closed vs Semantic Modules: Frequency

Number of Modules

Number of Axioms in a Module

CEL100
CEL250
CEL500
CEL750
CEL1000

MEX100
MEX250
MEX500
MEX750
MEX1000

Number of Modules

Number of Axioms in a Module
“I’ll be what you want me to be”

Consider again

\[ \text{Hindbrain\_hernia\_headache} \sqsubseteq \text{Headache\_disorder} \sqcap \exists \text{Due\_to. Cerebral\_herniation} \]
\[ \text{Removal\_f.b.\_from\_neck} \equiv \text{Neck\_operation} \sqcap \text{Removal\_foreign\_body} \]

But this time

\[ \Sigma = \{ \text{Hindbrain\_hernia\_headache}, \text{Removal\_f.b.\_from\_neck}, \text{Neck\_operation} \} \]

Now, \( M \) contains

\[ \text{Removal\_f.b.\_from\_neck} \equiv \]
\[ \text{Neck\_operation} \sqcap \text{Removal\_foreign\_body} \]

More flexibility in what does and what does not go into the module
Future Work

- **Logical diff**
  - Logical diff with a role box
    - Undecidable with arbitrary role inclusion axioms
    - $r \sqsubseteq s$ — no problem
    - Transitivity, left/right identities — under development
  - Logical diff for query answering
  - Logical diff based on model conservativity
  - More expressive languages
  - Pinpointing

- **Module extraction**
  - Minimal modules for $\mathcal{EL} +$ role box
  - Module extraction for more expressive languages
    - Non-tractable algorithms for minimal modules
    - **Locality** + minimal $\mathcal{EL}$-semantic modules