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).
- $\sim 400,000$ terms
- Simple structure:

  $$\text{Hindbrain\_hernia\_headache} \sqsubseteq \text{Headache\_disorder} \sqcap \exists \text{Due\_to.Cerebral\_herniation}$$

- $\mathcal{EL}$ representation

NB: $\mathcal{EL}$ will be a part of the updated OWL (??)

Understanding Changes in Terminologies

Suppose the following

$$\begin{aligned} \text{Neck\_injection} &\sqsubseteq \text{Operation} \\ \text{Neck\_operation} &\sqsubseteq \text{Operation} \\ \text{Removal\_f.b.\_from\_neck} &\equiv \text{Neck\_operation} \sqcap \text{Removal\_foreign\_body} \end{aligned}$$

is refined as

$$\begin{aligned} \text{Neck\_injection} &\sqsubseteq \text{Neck\_operation} \\ \text{Neck\_operation} &\sqsubseteq \text{Operation} \\ \text{Removal\_f.b.\_from\_neck} &\equiv \text{Neck\_operation} \sqcap \text{Removal\_foreign\_body} \end{aligned}$$

The refined terminology implies

- $\text{Neck\_injection} \sqsubseteq \text{Neck\_operation}$
- $\text{Neck\_injection} \sqcap \text{Removal\_foreign\_body} \sqsubseteq \text{Removal\_f.b.\_from\_neck}$

Logical Diff

What's the difference between different versions of a terminology?

- What's the difference over some signature Σ?

- $T_1,\ T_2$ – $\mathcal{EL}$ terminologies
- Σ – signature
- $\mathrm{diff}_\Sigma(T_1, T_2) = \left\{ C \sqsubseteq D \;\middle|\; \begin{array}{l} T_1 \not\models C \sqsubseteq D \\ T_2 \models C \sqsubseteq D \\ \mathrm{sig}(C \sqsubseteq D) \subseteq \Sigma \end{array} \right\}$

- We give a compact representation of $\mathrm{diff}_\Sigma(T_1, T_2)$
- Polytime algorithm

SNOMED CT'05 vs SNOMED CT'06: Joint Signature

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

SNOMED CT'05 vs SNOMED CT'06: Fragments

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

Size of Σ	CEX: diff(SNOMED CT'05,SNOMED CT'06)			
	Time (Sec.)	Memory (MByte)	$\lvert\mathsf{diffL}_\Sigma\rvert$	$\lvert\mathsf{diffR}_\Sigma\rvert$
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

- Role box ignored

Semantic Modules

- $\sim$ 10–15 minutes is fine, but we want better
- Time depends more on the terminology size than on Σ
- Extract modules!
  - Module should be a "replacement" for the terminology

Theorem

- *T — acyclic $\mathcal{EL}$ terminology*
- *No trivial axioms $A \equiv \top$, etc in T*

The *smallest semantic module can be computed in polytime*

CEX on MEX

Size of Σ	CEX: diff(SNOMED CT'05,SNOMED CT'06)				CEX: diff(Mod'05,Mod'06)	
	Time (Sec.)	Memory (MByte)	$\lvert\text{diffL}_\Sigma\rvert$	$\lvert\text{diffR}_\Sigma\rvert$	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

Minimal Semantic Modules

Consider

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

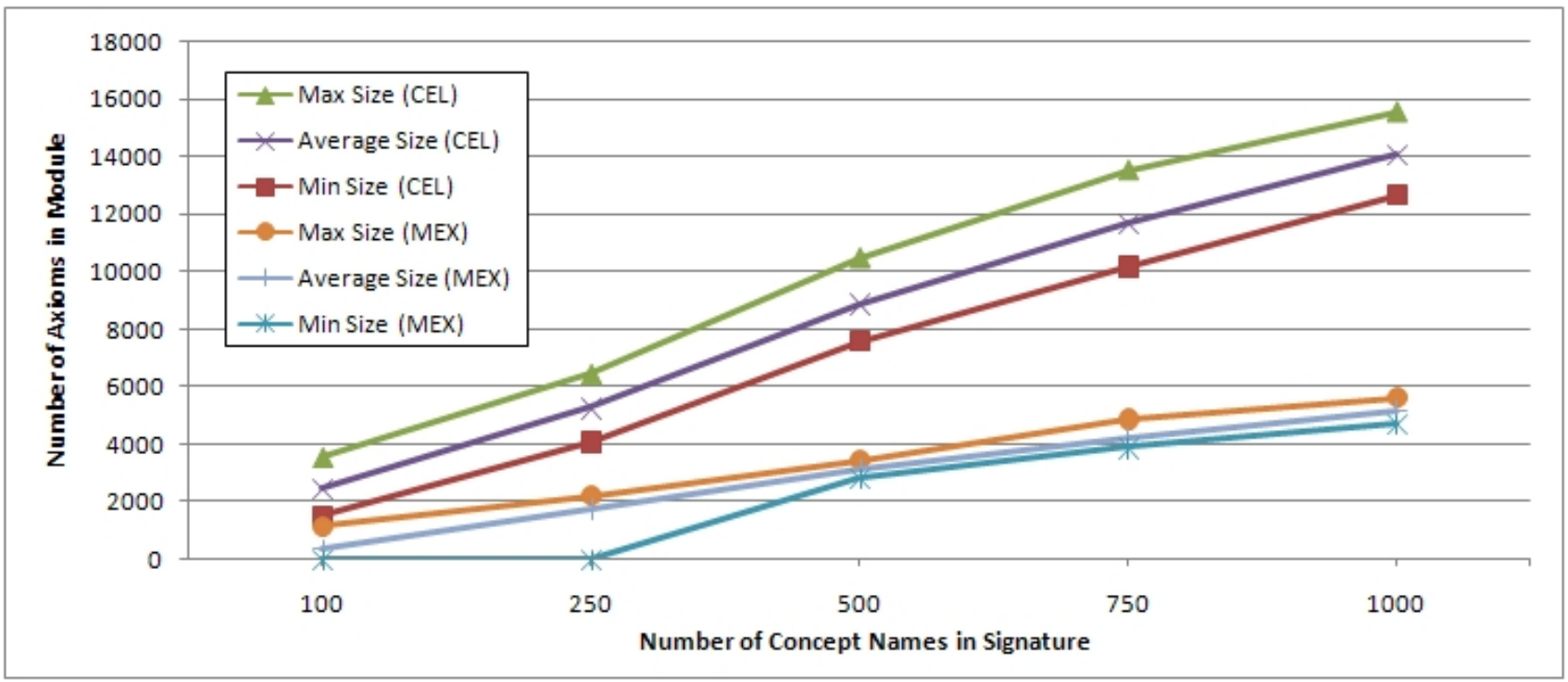
- $\Sigma = \{\text{Hindbrain\_hernia\_headache}, \text{Removal\_f.b.\_from\_neck}\}$
- $M = \emptyset$ (nothing!)

Many module extraction techniques would select both axioms!

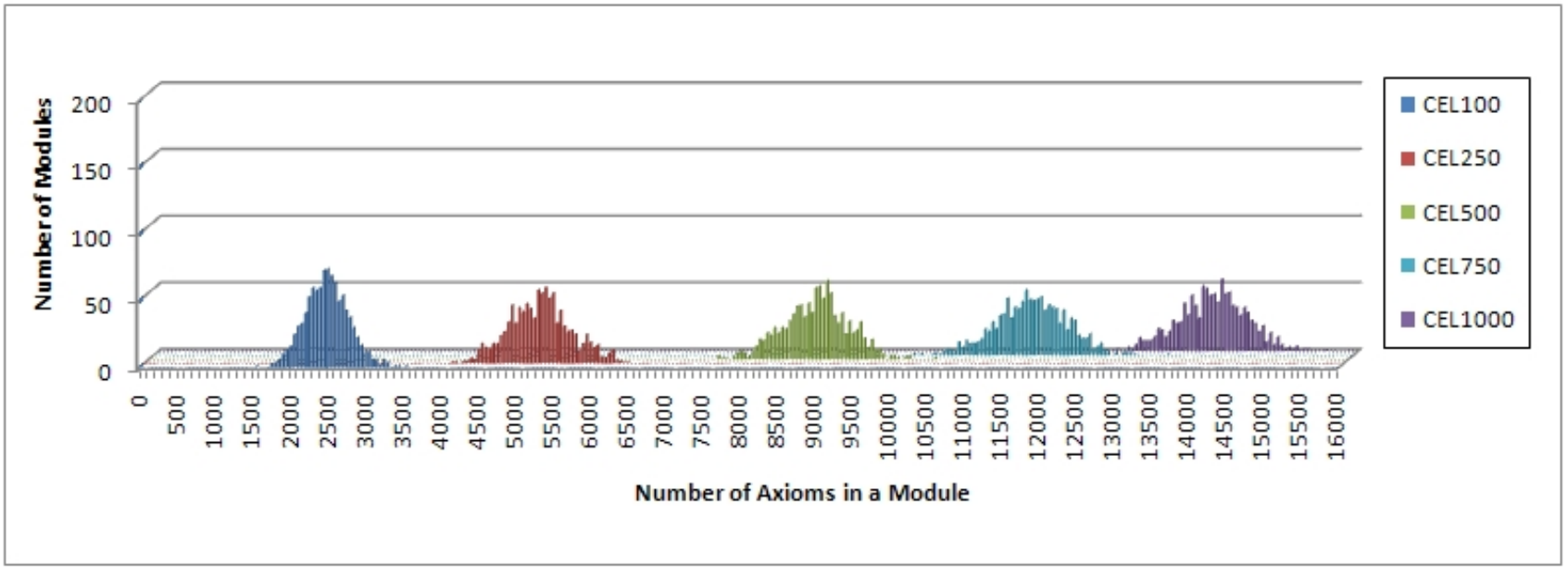
- definition closed

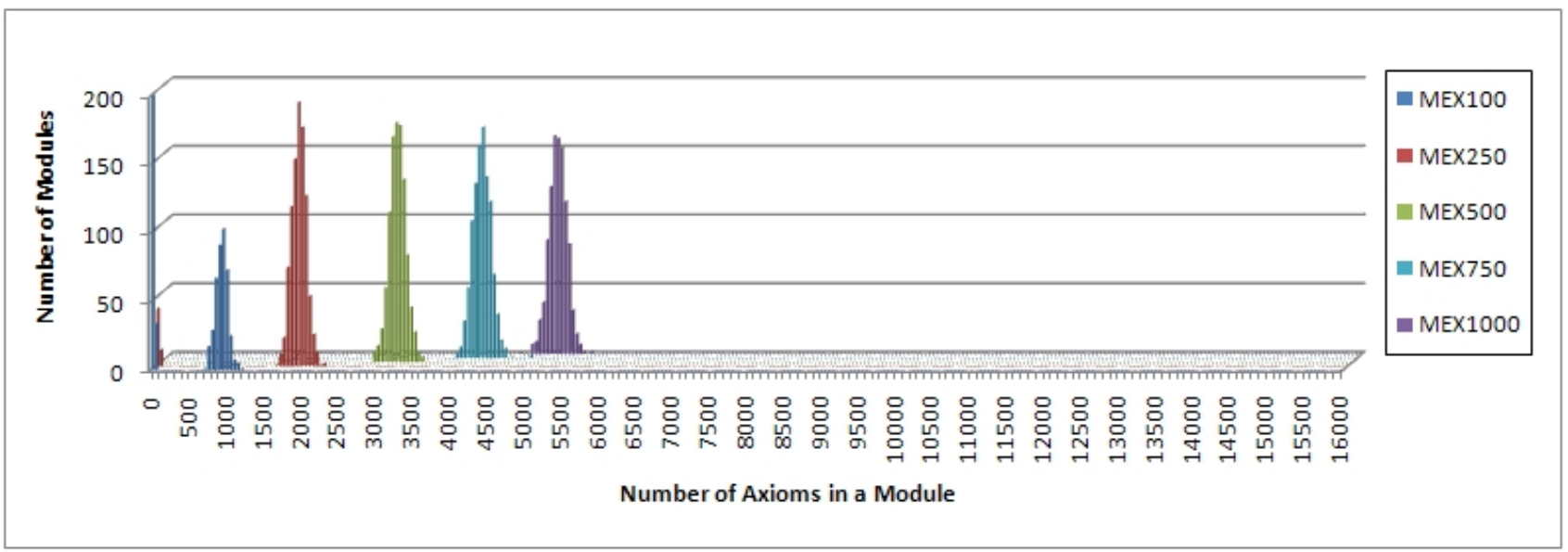
Definition-Closed vs Semantic Modules: Size

- Σ — randomly selected from SNOMED CT'05
- 1000 samples for each signature size
- with role box (under simplifying assumptions)



Definition-Closed vs Semantic Modules: Frequency





"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