Modularity in Ontologies: Comparison of module notions

Thomas Schneider[1] Dirk Walther[2]

[1]Department of Computer Science, University of Bremen, Germany

[2]Center for Advancing Electronics Dresden, TU Dresden, Germany

ESSLLI, 15 August 2013

Plan for this part

Comparison between

- MEX and locality-based modules
- modules based on syntactic and semantic locality

MEX experiments with SNOMED CT

SNOMED CT:

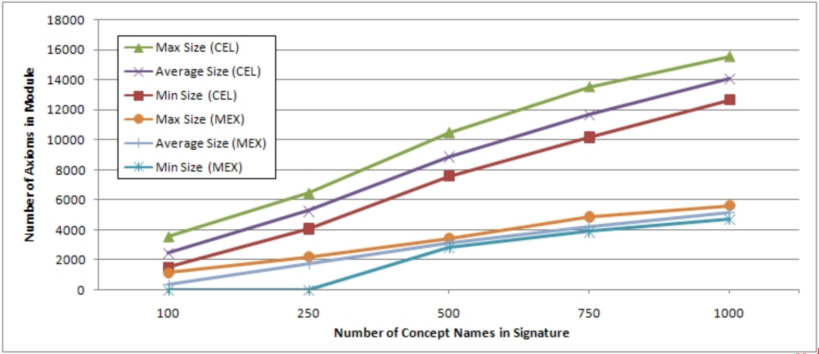
- Systematised Nomenclature of Medicine (Clinical Terms).
- $\sim$ 400,000 terms
- used in health care etc. in the US, UK, Australia etc.
- an **acyclic $\mathcal{EL}$-terminology** (+ role box):

[Konev, Lutz, Walther, Wolter 2008]
[Sattler, Schneider, Zakharyaschev 2009]

Experiment 1: Extraction of modules from SNOMED CT

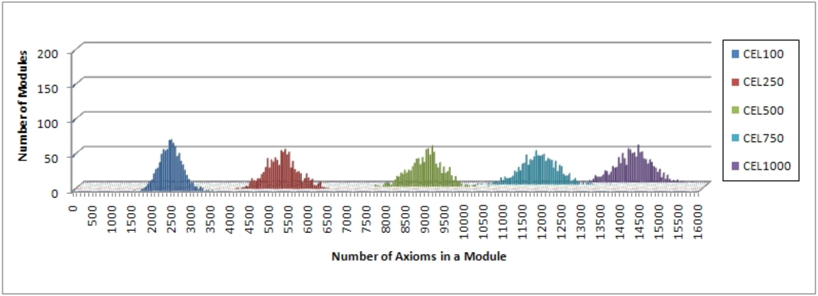
MEX: prototype implementation of the MEX algorithm[1]
vs. **CEL**: implementation of $\bot$-locality based modules

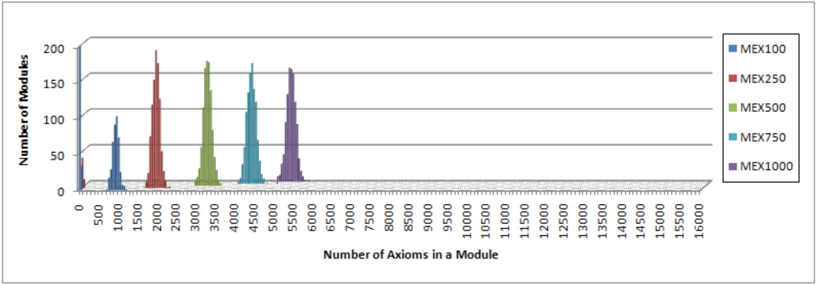
- Σ — randomly selected from SNOMED CT
- signature size up to 1000; for each size 1000 samples



[1] http://www.csc.liv.ac.uk/~konev/software/

MEX vs. ⊥-locality based modules: frequency

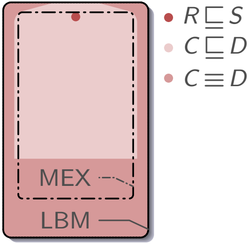




Experiment 2: SNOMED modules for clinical signatures

- Experiments with SNOMED again
- Compared modules for 24,000 terms from intensive care unit
- Locality-based modules (LBM) $\Leftrightarrow$ minimal modules (MEX)
- Results:

# axioms	
MEX	LBM
10%	15%
4–5 s	4–7 s



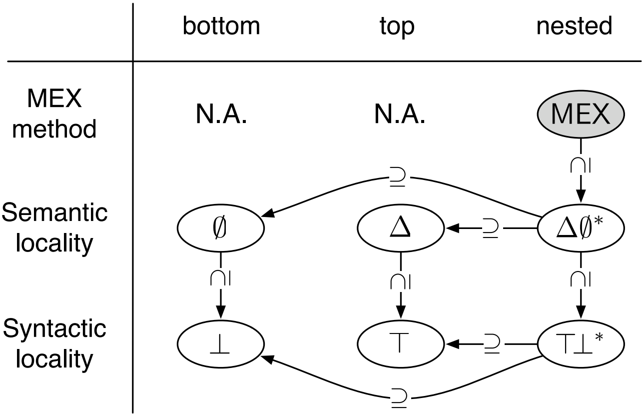
Preliminary conclusion

- MEX and locality-based modules are efficient to extract
- For random signatures from SNOMED,
  they differ significantly in size
- For clinical signatures from SNOMED, they don't differ much
- Most differences are caused by equivalence axioms
  (in fact, MEX = LBMs for equivalence-free $\mathcal{EL}$ terminologies)

Can this be generalised

- to other ontologies?
- to modules based on syntactic versus semantic locality?

Reminder: module notions



Questions

Given a **seed signature** Σ and ontology $\mathcal{O}$,

1. ... how likely is $\emptyset\text{-mod}(\Sigma, \mathcal{O}) \subset \bot\text{-mod}(\Sigma, \mathcal{O})$
   $\Delta\text{-mod}(\Sigma, \mathcal{O}) \subset \top\text{-mod}(\Sigma, \mathcal{O})$
   $\Delta\emptyset^*\text{-mod}(\Sigma, \mathcal{O}) \subset \top\bot^*\text{-mod}(\Sigma, \mathcal{O})$
   $\text{MEX-mod}(\Sigma, \mathcal{O}) \subset \Delta\emptyset^*\text{-mod}(\Sigma, \mathcal{O})$

   and how large is the difference?

   (**variation**: given axiom α,
   is it likely that α is $\emptyset$-local but not $\bot$-local for Σ
   Δ-local but not $\top$-local for Σ ?)

2. ... what is the difference in extraction time?

Sampling the seed signatures

- $\mathcal{O}$ has exponentially many potential seed signatures Σ.
- Modules for different Σ_1, Σ_2 may coincide.
- Still, $\mathcal{O}$ can have exponentially many modules.

  ➽ *Thursday* [Del Vescovo et al., 2010]
- We don't yet know what typical seed signatures are.

1. **Sample random seed signatures**
   - Sample one Σ: pick each axiom with probability $p = 0.5$
   - Achieve confidence interval $\pm 5\%$ with confidence level 95%: select 400 random Σ's (if $\mathcal{O}$ is big enough)

2. **Sample axiom seed signatures** (non-random, exhaustively)

   Genuine mod.s (GMs) ➽ *Thursday*

   - $\ldots$-mod(sig(α), $\mathcal{O}$), for $\alpha \in \mathcal{O}$
   - every module of $\mathcal{O}$ is the union of some GMs

The ontology corpus

Name	Expressivity	#Axioms	Sig. size
BioPortal (234 entries)	$\mathcal{AL}$–$\mathcal{SROIQ}(\mathcal{D})$	10–16,066	10–16,068
TONES			
Galen	$\mathcal{ALEHIF}+$	4,735	3,161
Koala	$\mathcal{ALCON}(\mathcal{D})$	42	32
Mereology	$\mathcal{SHIN}$	38	21
MiniTambis-rep'd	$\mathcal{ALCN}$	170	227
OWL-S Profile	$\mathcal{ALCHOIN}(\mathcal{D})$	276	163
People	$\mathcal{ALCHOIN}$	108	96
Tambis-full	$\mathcal{SHIN}(\mathcal{D})$	592	497
University	$\mathcal{SOIN}(\mathcal{D})$	52	44

Results: syntactic vs. semantic LBMs (1)

1. For 209 out of 242 ontologies,
   syntactic and semantic LBMs **agree**, i.e.:

   - Given an *arbitrary* Σ, there is **no difference** between

     - $\emptyset$-mod$(\Sigma, \mathcal{O})$ and $\bot$-mod$(\Sigma, \mathcal{O})$, or
     - Δ-mod$(\Sigma, \mathcal{O})$ and $\top$-mod$(\Sigma, \mathcal{O})$, or
     - $\Delta\emptyset^*$-mod$(\Sigma, \mathcal{O})$ and $\top\bot^*$-mod$(\Sigma, \mathcal{O})$, or
     - any α being $\emptyset$-local and $\bot$-local w.r.t. Σ, or
     - any α being Δ-local and $\top$-local w.r.t. Σ,

     at a significance level of 0.05.

   - Given *any* axiom signature sig(α), there is **no difference** between the syntactic and semantic LBM versions above

2. Extracting a $\emptyset$-module took up to $5\times$ as long as $\bot$-module (outlier: $34\times$ for Galen)

Results: syntactic vs. semantic LBMs (2)

For 6 of the remaining 33 ontologies, **negligible differences**:

- Differences are only caused by tautologies:
  - axioms like $\mathtt{r} \equiv (\mathtt{r}^-)^-$, for some role $\mathtt{r}$
  - contained in some BioPortal ontologies (published version is closed under certain entailments)
  - are not syntactically local for $r \in \Sigma$ but semantically local
  - sometimes "pull" other axioms into the module via signature extension
  - are uncritical: can be detected easily

$\leadsto$ **No observable differences** for 215 out of 242 ontologies

And the remaining 27?

Results: syntactic vs. semantic LBMs (3)

For the remaining 27 out of 242 ontologies,

- syntactic and semantic **modules differ** in only 6 cases
- differences between $\Delta\emptyset^*$-mod$(\Sigma, \mathcal{O})$ and $\top\bot^*$-mod$(\Sigma, \mathcal{O})$: at most 13 axioms
- larger differences only for Δ- vs. $\top$-modules
- time differences not measurable (few milliseconds per module)
- in the other 21 cases, only **locality of single axioms differs**

$\leadsto$ **Relevant module differences only in 6 of 242 ontologies!**

Differences are due to **3 patterns of axioms: culprits** (next)

One type of culprit

Example axiom α:

$$\mathtt{M} \equiv \underline{\mathtt{S}} \sqcap \forall \underline{\mathtt{c}}.\mathtt{F} \sqcap \forall \underline{\mathtt{g}}.\{\mathtt{m}\} \sqcap {=}3\, \underline{\mathtt{c}}.\top$$

```
EquivClasses(M,
S  and  c only F  and  g value m  and  c exactly 3 Thing)
```

- Suppose $\Sigma = \{\mathtt{S}, \mathtt{c}, \mathtt{g}\}$
- α is not $\bot$-local because none of its conjuncts is $\bot$-equiv.
- α is $\emptyset$-local:

  after replacing $\mathtt{M}, \mathtt{F}$ with $\bot$, it becomes a tautology
  in particular, $\forall \mathtt{c}.\bot \sqcap {=}3\, \mathtt{c}.\top$ cannot have any instances

LBMs vs. MEX: preprocessing

Q: How do LBMs compare with **minimal** modules?
$\leadsto$ Partial answer via MEX possible

Problem: MEX only defined for acyclic $\mathcal{ELI}$-TBoxes
So what can we do?

- Test only ontologies that comply?
  $\leadsto$ only 33 of 242 ☹
- Tweak + test ontologies that "almost" comply?
  $\leadsto$ only some 60 of 242 ☹
- Test $\mathcal{ELI}$-**approximation** of all ontologies! ☺

LBMs vs. MEX: $\mathcal{ELI}$-fication

Reduce every ontology to an acyclic $\mathcal{ELI}$ subset, removing

- all non-$\mathcal{ELI}$ axioms
- axioms involved in terminological cycles

This is a rather crude procedure.

Amount of reduction

- 33 ontologies are acyclic $\mathcal{ELI}$-terminologies
- from 36 ontologies, up to 28 axioms were removed
- from 170 ontologies, 30–12,185 axioms were removed

Compare LBMs and MEX for this new corpus

LBMs vs. MEX: result overview

- Diffs MEX–LBMs in $\sim$ **27%** of the preprocessed ontologies
- for these, no diffs syntactic–semantic LBM

Experiment	#ontol. with diffs.	% tests with diffs.	avg size of diffs #axs	rel.
Random signatures	66	84%	0–26	0–13%
Axiom signatures	61	12%	0–13	0–80%

- Largest differences: Galen with 127 axioms (outlier)
- same differences occur for many seed signatures
  $\rightsquigarrow$ probably caused by features of the ontology

Q: Do the differences correlate with ontology size, expressivity, or amount of modification ($\mathcal{ELI}$-fication)?

LBMs vs. MEX: results by ontology measures

Group	#axioms removed	#ontologies	ontology size (avg.)
1 unchanged ontologies *no* diff. $\Delta\emptyset^* \setminus$ MEX	0	33 (14%)	19–16,066 (2,176)
2 little-changed ontologies *no* diff. $\Delta\emptyset^* \setminus$ MEX	1–28	36 (15%)	13– 6,587 (466)
3 largely-changed ontologies *no* diff. $\Delta\emptyset^* \setminus$ MEX	31–7,836 (avg. 884)	104 (44%)	51–13,153 (2,373)
4 largely-changed ontologies *with* diff. $\Delta\emptyset^* \setminus$ MEX	30–12,185 (avg. 1,001)	66 (27%)	42–12,344 (1,843)

Differences correlate with

- expressivity
  (Group 1+2 mostly $\mathcal{EL}$; Group 4 highly expressive, e.g., nominals)
- amount of $\mathcal{ELI}$-fication
  (only "largely-changed" ontologies show differences)
- *not* with size

Culprits: equivalence axioms $A \equiv C$

Summary of module comparison

- Only 6 out of 242 ontologies showed non-trivial differences between semantic and syntactic LBMs
- These differences are small
- Theoretically hard semantic LBMs are often easy to compute

- Only 66 out of 242 $\mathcal{ELI}$-fied ontologies showed differences between LBMs and MEX
- Many of these differences are rather small

$\leadsto$ **Cheap is cheerful!**