Syntactic vs. semantic locality

How good is a cheap approximation?

Chiara Del Vescovo\textsuperscript{1}  Pavel Klinov\textsuperscript{2}  Bijan Parsia\textsuperscript{1}
Uli Sattler\textsuperscript{1}  \textit{Thomas Schneider}\textsuperscript{3}  Dmitry Tsarkov\textsuperscript{1}

\textsuperscript{1}School of Computer Science, University of Manchester, UK
\textsuperscript{2}Institute of Artificial Intelligence, University of Ulm, Germany
\textsuperscript{3}Dept. of Computer Science, University of Bremen, Germany

WoMO  Graz, 24 July 2012
Overview

1 Syntactic vs. semantic locality

2 How good is a cheap approximation?
And now . . .

1. Syntactic vs. semantic locality

2. How good is a cheap approximation?
A reuse scenario

“Borrow” knowledge from external ontologies

- Provides access to well-established knowledge
- Doesn’t require expertise in external disciplines

This scenario is well-understood and implemented.
A reuse scenario

Reuse one external, *monolithic* ontology
A reuse scenario

Reuse one external, *monolithic* ontology

- **NCI**
- **knowledge about “Disease” and “Arthritis”**
- **JRA Ontology**
A reuse scenario

Reuse one external, *monolithic* ontology

![Diagram showing NCI and JRA Ontology with knowledge about "Disease" and "Arthritis"]

How much of NCI do we need?
A reuse scenario

Reuse *a part* of an external, monolithic ontology

How much of **NCI** do we need?

- **Coverage:** Import *everything* relevant for the chosen terms.
- **Economy:** Import *only* what’s relevant for them. Compute that part quickly.
A reuse scenario

Reuse parts of several external ontologies

- NCI
  - Disease, Arthritis
- Galen
  - Drug, Joint
- JRA Ontology
A reuse scenario

Syntactic vs. semantic locality

How good is a cheap approximation?

NCI

JRAO

Galen

Arthritis diseases

Arthropathy

Arthritis

Autoimmune Disease

Rheumatologic Disorder

Atrophic Arthritis

Polyarthritis

Rheumatoid Arthritis

Juvenile Chronic Polyarthritis

Juvenile Rheumatoid Arthritis

C1

C7

Drugs

Joints

affects

isTreatedBy

Thomas Schneider

Syntactic vs. semantic locality
What is a module?

\( \mathcal{M} \) is a *module* of \( \mathcal{O} \) for signature \( \Sigma \):

- \( \mathcal{M} \subseteq \mathcal{O} \)
- \( \mathcal{M} \) covers \( \mathcal{O} \) for \( \Sigma \), i.e.,

  for all compatible \( \mathcal{O}' \),
  \( \mathcal{O}' \cup \mathcal{M} \) preserves all knowledge about \( \Sigma \) in \( \mathcal{O}' \cup \mathcal{O} \).
What is a module?

\[ M \text{ is a } \text{module} \text{ of } O \text{ for signature } \Sigma: \]

- \[ M \subseteq O \]
- \[ M \text{ covers } O \text{ for } \Sigma, \text{ i.e.,} \]
  
  for all \( O' \) that share only \( \Sigma \)-terms with \( O \),
  
  for all axioms \( \eta \) built from terms in \( \Sigma \):
  
  if \( \eta \) follows from \( O' \cup O \), then \( \eta \) follows from \( O' \cup M \).
What is a module?

\( M \) is a *module* of \( O \) for signature \( \Sigma \):
- \( M \subseteq O \)
- \( M \) covers \( O \) for \( \Sigma \), i.e.,
  - for all compatible \( O' \),
    \( O' \cup M \) preserves all knowledge about \( \Sigma \) in \( O' \cup O \).
What is a module?

\( \mathcal{M} \) is a *module* of \( \mathcal{O} \) for signature \( \Sigma \):

- \( \mathcal{M} \subseteq \mathcal{O} \)
- \( \mathcal{M} \) covers \( \mathcal{O} \) for \( \Sigma \), i.e.,
  
  for all compatible \( \mathcal{O}' \),
  
  \( \mathcal{O}' \cup \mathcal{M} \) preserves all knowledge about \( \Sigma \) in \( \mathcal{O}' \cup \mathcal{O} \).

- Coverage \( \models \) preserving entailments;
  
  Without coverage: no encapsulation \( \not\sim \) no module

**Fact:** \( \mathcal{M} \) covers \( \mathcal{O} \) for \( \Sigma \) iff \( \mathcal{O} \) is a \( \Sigma \)-CE of \( \mathcal{M} \);

\( \mathcal{O}' \) doesn't determine what counts as a module.
What is a module?

\( \mathcal{M} \) is a module of \( \mathcal{O} \) for signature \( \Sigma \):

- \( \mathcal{M} \subseteq \mathcal{O} \)
- \( \mathcal{M} \) covers \( \mathcal{O} \) for \( \Sigma \), i.e.,

  for all compatible \( \mathcal{O'} \),
  \( \mathcal{O'} \cup \mathcal{M} \) preserves all knowledge about \( \Sigma \) in \( \mathcal{O'} \cup \mathcal{O} \).

Coverage \( \hat{=} \) preserving entailments;

Without coverage: no encapsulation \( \leadsto \) no module

- \( \mathcal{O'} \cup \mathcal{O} \) is called \( \Sigma\text{-conservative extension (CE)} \) of \( \mathcal{O'} \cup \mathcal{M} \)
  
  [Ghilardi, Lutz, Wolter 2006]

- Fact: \( \mathcal{M} \) covers \( \mathcal{O} \) for \( \Sigma \) iff \( \mathcal{O} \) is a \( \Sigma\text{-CE} \) of \( \mathcal{M} \)

  \( \leadsto \) \( \mathcal{O'} \) doesn’t determine what counts as a module
How is a minimal $\Sigma$-module extracted?

Simple module extraction algorithm:

- $\mathcal{M} \leftarrow \mathcal{O}$
- While $\mathcal{M} \setminus \{\alpha\}$ covers $\mathcal{O}$ for $\Sigma$, for some $\alpha \in \mathcal{M}$, remove $\alpha$ from $\mathcal{M}$.
- Output $\mathcal{M}$
How is a minimal $\Sigma$-module extracted?

Simple module extraction algorithm:

- $M \leftarrow O$
- While $M \setminus \{\alpha\}$ covers $O$ for $\Sigma$, for some $\alpha \in M$, remove $\alpha$ from $M$.
- Output $M$

Observation:
Different orders of choosing $\alpha$
can lead to different minimal modules
How is a minimal $\Sigma$-module extracted?

Simple module extraction algorithm:

1. $M \leftarrow O$
2. While $M \setminus \{\alpha\}$ covers $O$ for $\Sigma$, for some $\alpha \in M$, remove $\alpha$ from $M$.
3. Output $M$

Problem:

How to decide the CE property?

Usually harder than standard reasoning, often undecidable!

Approximation 1: semantic locality \((\emptyset, \Delta)\)

\(\mathcal{M}\) is a \(\emptyset\)-module of \(\mathcal{O}\) for \(\Sigma\):

if every ax. \(\alpha\) in \(\mathcal{O} \setminus \mathcal{M}\) is \(\emptyset\)-local for \(\Sigma\)

i.e., if all non-\(\Sigma\) symbols are replaced by \(\bot\),
then \(\alpha\) becomes a tautology

[Cuenca Grau et al. 2007]
Approximation 1: semantic locality $(\emptyset, \Delta)$

$\mathcal{M}$ is a $\emptyset$-module of $\mathcal{O}$ for $\Sigma$:

- if every ax. $\alpha$ in $\mathcal{O} \setminus \mathcal{M}$ is $\emptyset$-local for $\Sigma$

  i.e., if all non-$\Sigma$ symbols are replaced by $\bot$, then $\alpha$ becomes a tautology

  [Cuenca Grau et al. 2007]

Facts:

- $\emptyset$-module of $\mathcal{O}$ for $\Sigma$ is uniquely determined
- *the* $\emptyset$-$\text{mod}(\Sigma, \mathcal{O})$ covers $\mathcal{O}$ for $\Sigma$
  - (but $\mathcal{M}$ isn’t necessarily a *minimal* module)
- Deciding semantic locality is as hard as reasoning
Approximation 1: semantic locality \((\emptyset, \Delta)\)

\(M\) is a \(\emptyset\)-module of \(O\) for \(\Sigma\):

- if every ax. \(\alpha\) in \(O \setminus M\) is \(\emptyset\)-local for \(\Sigma\)
- i.e., if all non-\(\Sigma\) symbols are replaced by \(\perp\), then \(\alpha\) becomes a tautology

[Cuenca Grau et al. 2007]

Facts:

- \(\emptyset\)-module of \(O\) for \(\Sigma\) is uniquely determined
- \(\text{the } \emptyset\text{-mod}(\Sigma, O)\) covers \(O\) for \(\Sigma\)
  - (but \(M\) isn’t necessarily a minimal module)
- Deciding semantic locality is as hard as reasoning

Dual notion: \(\Delta\)-module, \(\Delta\)-locality
Approximation 2: syntactic locality ($\perp, T$)

$\mathcal{M}$ is a $\perp$-module of $\mathcal{O}$ for $\Sigma$:

if every ax. $\alpha$ in $\mathcal{O} \setminus \mathcal{M}$ is $\perp$-local for $\Sigma$

i.e., $\alpha$ is generated by a grammar
that describes obviously $\emptyset$-local axioms
for the DL $\mathcal{SROIQ}$ underlying OWL 2  [Cuenca Grau et al. 2007]
Approximation 2: syntactic locality \((\bot, \top)\)

\(\mathcal{M}\) is a \(\bot\)-module of \(\mathcal{O}\) for \(\Sigma\):

- if every ax. \(\alpha\) in \(\mathcal{O} \setminus \mathcal{M}\) is \(\bot\)-local for \(\Sigma\)

i.e., \(\alpha\) is generated by a grammar that describes obviously \(\emptyset\)-local axioms
for the DL \(\text{SROIQ}\) underlying \(\text{OWL 2}\) [Cuenca Grau et al. 2007]

Facts:

- \(\text{The } \bot\text{-mod}(\Sigma, \mathcal{O})\) contains the \(\emptyset\text{-mod}(\Sigma, \mathcal{O})\)
  and hence covers \(\mathcal{O}\) for \(\Sigma\)
  (but again isn’t necessarily a minimal module)

- syntactic locality can be decided efficiently: in poly-time!
Approximation 2: syntactic locality $(\bot, \top)$

$\mathcal{M}$ is a $\bot$-module of $\mathcal{O}$ for $\Sigma$:

- if every ax. $\alpha$ in $\mathcal{O} \setminus \mathcal{M}$ is $\bot$-local for $\Sigma$ 

i.e., $\alpha$ is generated by a grammar that describes obviously $\emptyset$-local axioms for the DL $SROIQ$ underlying OWL 2 [Cuenca Grau et al. 2007]

Facts:

- The $\bot$-mod($\Sigma$, $\mathcal{O}$) contains the $\emptyset$-mod($\Sigma$, $\mathcal{O}$) and hence covers $\mathcal{O}$ for $\Sigma$

  (but again isn’t necessarily a minimal module)

- syntactic locality can be decided efficiently: in poly-time!

Dual notion: $\top$-module, $\top$-locality
Summary: locality-based modules (LBMs)

- Syntactic LBMs are cheap for DLs up to OWL
- Semantic LBMs are expensive for expressive DLs
  
  (and infeasible for FOL)

- All LBMs provide coverage, but do not guarantee minimality

- Conservativity-based modules are infeasible for expressive DLs and FOL
And now . . .

1. Syntactic vs. semantic locality

2. How good is a cheap approximation?
Facts about syntactic locality based modules (LBMs)

- $\bot$-mod and $\top$-mod have been implemented: OWL API etc.
- More economic: $\top\bot^*$-mod (alternative nesting until fixpoint)
- Previous experiments: $\top\bot^*$-mod often well-sized in practice
  - Experiments with SNOMED ($\mathcal{EL}$, 350,000 axioms)
  - Compared modules for 24,000 terms from intensive care unit
  - $\top\bot^*$-mod (LBM) $\Leftrightarrow$ module based on model-CE (MEX)
Facts about *syntactic* locality based modules (LBM)

- $\bot$-mod and $\top$-mod have been implemented: OWL API etc.
- More economic: $\top\bot^*$-mod (alternative nesting until fixpoint)
- Previous experiments: $\top\bot^*$-mod often well-sized in practice
  - Experiments with SNOMED ($\mathcal{EL}$, 350,000 axioms)
  - Compared modules for 24,000 terms from intensive care unit
  - $\top\bot^*$-mod (LBM) $\Leftrightarrow$ module based on model-CE (MEX)

<table>
<thead>
<tr>
<th></th>
<th>MEX</th>
<th>LBM</th>
</tr>
</thead>
<tbody>
<tr>
<td># axioms</td>
<td>10%</td>
<td>15%</td>
</tr>
<tr>
<td>4–5 s</td>
<td>4–7 s</td>
<td></td>
</tr>
</tbody>
</table>

Results:

- $R \sqsubseteq S$
- $C \sqsubseteq D$
- $C \equiv D$
Can you take a little off, please?

*Semantic* LBMs ($\emptyset$-mod and $\Delta$-mod)

- ... are contained in the respective syntactic LBM, remember:

\[
\emptyset\text{-mod}(\Sigma, O) \subseteq \bot\text{-mod}(\Sigma, O) \\
\Delta\text{-mod}(\Sigma, O) \subseteq T\text{-mod}(\Sigma, O)
\]

- ... are extracted using reasoning

- ... have not been implemented yet

Are they actually (typically, significantly, ...) smaller?
How much more expensive is their extraction?
Questions

Given a signature $\Sigma$ and ontology $\mathcal{O}$,

1. ... how likely is $\emptyset\text{-mod}(\Sigma, \mathcal{O}) \subset \bot\text{-mod}(\Sigma, \mathcal{O})$, and how large is the difference?

   (variation: given axiom $\alpha$, is it likely that $\alpha$ is $\emptyset$-local but not $\bot$-local for $\Sigma$?)

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

Given a signature $\Sigma$ and ontology $\mathcal{O}$,

1. . . how likely is $\emptyset\text{-mod}(\Sigma, \mathcal{O}) \subset \perp\text{-mod}(\Sigma, \mathcal{O})$, and how large is the difference?

   (variation: given axiom $\alpha$, is it likely that $\alpha$ is $\emptyset$-local but not $\perp$-local for $\Sigma$?)

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

Later: the same questions for the pairs

- $\Delta\text{-mod}$ vs. $\top\text{-mod}$
- $\Delta\emptyset^*\text{-mod}$ vs. $\top\perp^*\text{-mod}$
Sampling the seed signatures

- $\mathcal{O}$ has exponentially many potential seed signatures $\Sigma$.
- Modules for different $\Sigma_1, \Sigma_2$ may coincide.
- Still, $\mathcal{O}$ can have exp. many modules. [Del Vescovo et al., 2010]
- We don’t yet know what typical seed signatures are.
Sampling the seed signatures

- $\mathcal{O}$ has exponentially many potential seed signatures $\Sigma$.
- Modules for different $\Sigma_1, \Sigma_2$ may coincide.
- Still, $\mathcal{O}$ can have exp. many modules. [Del Vescovo et al., 2010]
- We don’t yet know what typical seed signatures are.

Sample random seed signatures!

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

- $\mathcal{O}$ has exponentially many potential seed signatures $\Sigma$.
- Modules for different $\Sigma_1, \Sigma_2$ may coincide.
- Still, $\mathcal{O}$ can have exp. many modules. [Del Vescovo et al., 2010]
- We don’t yet know what typical seed signatures are.

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

- Non-random seed signatures
  
  *Genuine mod.s (GMs)*
  - $\perp$-mod($\text{Sig}(\alpha), \mathcal{O}$), for $\alpha \in \mathcal{O}$
  - every module of $\mathcal{O}$ is the union of some GMs

  $\leadsto$ include axiom signatures $\text{Sig}(\alpha)$
## The ontology corpus

<table>
<thead>
<tr>
<th>Name</th>
<th>Expressivity</th>
<th>#Axioms</th>
<th>Sig. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioPortal (148 entries)</td>
<td>$ALCN \rightarrow SHIN(\mathcal{D})/SOIN(\mathcal{D})$</td>
<td>38–4,735</td>
<td>21–3,161</td>
</tr>
<tr>
<td>TONES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galen</td>
<td>$ALEHIF+$</td>
<td>4,735</td>
<td>3,161</td>
</tr>
<tr>
<td>Koala</td>
<td>$ALCON(\mathcal{D})$</td>
<td>42</td>
<td>32</td>
</tr>
<tr>
<td>Mereology</td>
<td>$SHIN$</td>
<td>38</td>
<td>21</td>
</tr>
<tr>
<td>MiniTambis-rep’d</td>
<td>$ALCN$</td>
<td>170</td>
<td>227</td>
</tr>
<tr>
<td>OWL-S Profile</td>
<td>$ALCHOIN(\mathcal{D})$</td>
<td>276</td>
<td>163</td>
</tr>
<tr>
<td>People</td>
<td>$ALCHOIN$</td>
<td>108</td>
<td>96</td>
</tr>
<tr>
<td>Tambis-full</td>
<td>$SHIN(\mathcal{D})$</td>
<td>592</td>
<td>497</td>
</tr>
<tr>
<td>University</td>
<td>$SOIN(\mathcal{D})$</td>
<td>52</td>
<td>44</td>
</tr>
</tbody>
</table>
Results I: cheap is good!

1. For 151 out of 156 ontologies, $\emptyset$-mod and $\perp$-mod agree, i.e.:
   - Given an arbitrary $\Sigma$, there is no difference between
     - $\emptyset$-mod($\Sigma, O$) and $\perp$-mod($\Sigma, O$), and
     - any $\alpha$ being $\emptyset$-local and $\perp$-local w.r.t. $\Sigma$,
   at a significance level of 0.05.
   - Given any axiom signature $\text{Sig}(\alpha)$, there is no difference between $\emptyset$-mod($\text{Sig}(\alpha), O$) and $\perp$-mod($\text{Sig}(\alpha), O$).

2. Extracting a $\emptyset$-module took up to $6 \times$ as long as $\perp$-module (average $2.7 \times$)
Results II: cheap seems good enough

1. For 2 ontologies from BioPortal,\(^1\) negligible differences:
   - Up to 30 out of 3,446 (resp. 6,008) axioms
   - Axioms are: \( r \equiv (r^-)^- \), for some role (object property) \( r \)
     i.e., \( \text{EquivObjProps}(r, \text{inv(inv}(r))) \)
   - Uncritical: these are few tautologies
     (Published version of some BioPortal ontologies is closed under certain entailments)

2. Extraction time up to 6\(\times\) on average

---

\(^1\)Experimental Factor Ontology and Software Ontology

Thomas Schneider
Results III: a single type of culprit

For the remaining 3 ont.s,\(^2\) small differences of 1 common pattern

Example axiom \(\alpha\): 

\[
M \equiv S \cap \forall c. F \cap \forall g. \{m\} \cap = 3 \ c. \top
\]

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

\(^2\)Koala, miniTambis and Tambis
Results III: a single type of culprit

For the remaining 3 ont.s, small differences of 1 common pattern

Example axiom $\alpha$:

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

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

- Suppose $\Sigma = \{ S, c, g \}$
- $\alpha$ is not $\bot$-local because none of its conjuncts is $\bot$-equiv.
- $\alpha$ is $\emptyset$-local:
  after replacing $M, F$ with $\bot$, it becomes a tautology
  in particular, $\forall c. \bot \sqcap = 3 c. \top$ cannot have any instances

---

$^2$Koala, miniTambis and Tambis
Results IV: cheap still seems good enough

1. These culprits have
   - no effects on Koala modules
     (only singleton differences for locality)
   - small effects on miniTambis:
     - $\bot$-modules up to 4 axioms (3%) larger than $\emptyset$-modules
     - $\bot$-GMs up to 7 axioms (75%) larger than $\emptyset$-GMs
   - small effects on Tambis:
     - $\bot$-modules up to 11 axioms (2%) larger than $\emptyset$-modules
     - $\bot$-GMs up to 41 axioms (26%) larger than $\emptyset$-GMs

2. Extraction time up to $5\times$ on average for Tambis
   (not measurable for Koala and miniTambis)
Δ-modules cannot always be extracted using DL reasoners:

- Remember – locality check: replace non-Σ symbols with $\top$ and test for tautology
- Global restrictions of $SROIQ$ don’t allow $\top$-role in number restrictions or role chains
- This affects 39 ontologies in our corpus

For the remaining 117 ontologies, there is no (statistically significant) difference:

- between Δ- and $\top$-modules
- between $\Delta\varnothing^*$- and $\top\bot^*$-modules
Lessons learnt

- No or little difference between semantic and syntactic locality
- Syntactic locality seems a good approximation of semantic locality
- Cheap is good!
  - (Still, semantic module extraction often fast in practice)
Outlook

- Incorporate the missing 39 (richer) ontologies into
  - $\Delta$- vs. $\top$-locality
  - $\Delta\emptyset^*$- vs. $\top\bot^*$-locality

- Extend study to larger ontologies
  - NCI has axioms that nest the culprit pattern
  - Not reproducible with the official releases

- Modify sampling
  - Put more weight on small and large seed signatures
  - Measure difference w.r.t. a given module
    - $\sim$ sampling of modules instead of seed signatures

- Include conservativity-based modules (for lightweight DLs)
Incorporate the missing 39 (richer) ontologies into
- $\Delta$- vs. $\top$-locality
- $\Delta\emptyset^*$- vs. $\top\perp^*$-locality

Extend study to larger ontologies
- NCI has axioms that nest the culprit pattern
- Not reproducible with the official releases

Modify sampling
- Put more weight on small and large seed signatures
- Measure difference w.r.t. a given module
  $\rightsquigarrow$ sampling of modules instead of seed signatures

Include conservativity-based modules (for lightweight DLs)

Thank you.