Modularity Meets Forgetting: A Case Study with the SNOMED CT Ontology (Extended Abstract)

Jieying Chen  Ghadah Alghamdi  Renate A. Schmidt
The University of Manchester*
Manchester, UK

Dirk Walther  Yongsheng Gao
DNV GL  IHTSDO
Oslo, Norway  London, UK

SNOMED CT\[1\] is a comprehensive, precise and widely-used medical ontology covering ample clinical specialities and requirements. The latest release from January 2019 contains more than 340 000 axioms, with the number of axioms increasing by about 10% compared to the version of 2016. The number of concept names and size of the ontology will keep on increasing further.

Maintaining and developing such a large-scale ontology poses significant challenges for semantic web applications. Ontology modularization and forgetting tools provide automated support for ontology development for extracting smaller sets of relevant axioms from an ontology. In this presentation we discuss the theoretical relationship of three existing module extraction techniques and investigate their practical benefits when combined with forgetting.

Creation of ontology extracts is a useful operation in the reuse, creation, curation, decomposition, integration and general use of ontologies. For example, for reviewing and analyzing the information relating to the concept “kidney disease (disorder)” which has more than 1200 sub-concepts in SNOMED CT, developers would benefit from being able to work with an extract that succinctly summarizes all information in the ontology relating to kidney diseases. Another concrete scenario is a doctor wishing to find diseases that have an inflammatory morphology and a finding site of kidney structure based on morphologies and/or finding sites. Instead of querying SNOMED CT as a whole, it would be more efficient to simply query a smaller extract of the ontology containing sufficiently many axioms to compute the same answer as if it was computed on the entire ontology.

A method commonly used for creating extracts of an ontology is modularization. In general, a module of an ontology is a subset of the ontology that in a specific context can function in the same way as the original ontology. Different notions of modules have been proposed for different specific functions. Among them are notions such as so-called plain, self-contained and depleting modules \[6\]. The MEX system extracts minimal depleting and self-contained modules from ontologies formulated as acyclic $\mathcal{EL}$-terminologies \[1\] that allows conjunction and existential restriction to construct concept expressions. The recently introduced minimal subsumption modules \[3, 4\] are subsets of an ontology preserving $\mathcal{EL}$-subsumption queries. The evaluation in \[4\] shows that minimal subsumption modules are generally much smaller than MEX-modules. However, deciding the preservation of subsumption queries can be expensive, and the algorithm for computing minimal subsumption modules from $\mathcal{EL}$-terminologies runs in exponential time.

\*This note describes ongoing work undertaken in EPSRC IAA Project 228 on “Comparison and Abstraction of SNOMED CT Ontologies”. Initial findings were presented at the 32nd International Workshop on Description Logics \[2\]. We thank Yizheng Zhao for most valuable help with the FAME system.

[1] https://www.snomed.org

© J.Chen, G.Alghamdi, R.Schmidt, D.Walther, Y.Gao
This work is licensed under the Creative Commons Attribution License.
Approximate modules, such as modules based on syntactic locality \cite{5}, can be computed more efficiently. For instance, the algorithm of extracting locality-based modules runs in polynomial time in the size of the ontology. Empirical investigations by \cite{5,6} in application-close scenarios involving SNOMED CT have found that, while graph-based approaches to modularization have reasonable coverage (71%-96%), the obtained extracts are large (17%-51% of the size of SNOMED CT) \cite{8}. While relatively small extracts can be obtained with locality-based modules, a down-side is lower precision due to the presence typically of a large number of symbols in the module outside the desired signature specified as input.

An alternative method for creating a compact representation of a subset of the information in an ontology is (deductive) forgetting, also known as uniform interpolation. Forgetting allows the creation of abstractions of an ontology by hiding specified concept or role names (the forgetting signature), without losing the underlying logical definitions of the remaining concept and role names (the interpolation signature). Because deductive forgetting preserves all information in the specified interpolation signature it has high precision, but the axioms in a forgetting solution (or uniform interpolant) are in general not axioms belonging to the original ontology. For example, interpolating the ontology

\[ \varnothing = \{ A \sqsubseteq \exists r. C, C \sqsubseteq B \} \]

for the concept names \( A \) and \( B \) (or forgetting \( C \)) is \( \{ A \sqsubseteq \exists r. B \} \), while the module for \( A \) and \( B \) is the ontology itself. Theoretical investigations of uniform interpolation and forgetting for ontologies include \cite{9,10}.

Compared with ontology modularization, a significant advantage of deductive forgetting is that the returned forgetting solutions (interpolants) only use class and property names in the interpolation signature.

Our interest in this paper is the computation of forgetting-based ontology extracts in real scenarios involving the SNOMED CT ontology. Such scenarios require nearly all of the names in the ontology to be forgotten, which poses a significant challenge to forgetting tools. In particular, role names can be difficult to forget. The approach introduced in this paper uses a pipeline of three steps:

1. extension, and if needed partitioning, of the given interpolation signature,
2. modularization, in particular, we evaluate three different forms of modularization: locality-based star-modules, modularization performed by the MEX system and minimal subsumption modularization, and
3. forgetting using the systems FAME \cite{11} and LETHE \cite{7}. LETHE implements a resolution-based approach to forgetting, while FAME implements a hybrid that incorporates steps from Ackermann-based forgetting.

The results show dividing the problem based on signature partitioning and precomputed ontology modules can help to reduce the number of symbols that need to be forgotten. Minimal subsumption modules in particular have been found to significantly reduce the size of extracts. An advantage of our approach is the 100% precision of the returned views as forgetting solutions, which cannot be achieved with other approaches.

The paper addresses the following issues posed by the creation of extracts of ontologies based on modularization and forgetting.

1. Interpolation signatures specified as topics in SNOMED CT defined by a refset of concepts. Refsets are lists of concept names that have specific meaning, such as a list of kidney diseases.
2. The given interpolation signature is very small in the size of the signature of SNOMED CT and the forgetting signature is correspondingly very large. (The average size of NHS Refsets is less than 0.8% of the signatures in SNOMED CT and ERA Refset is only 0.02%.)

3. Extending the concept set to include roles and concepts picked up ‘horizontally’, thereby creating a neighbourhood of related concept and role names for the given refset.

4. Since locality-based modules are often large (unless the seed signature is very carefully chosen), and usually contain foreign symbols, and for the other forms of modularization producing smaller results the forgetting tools still had difficulties, a key issue was creating modules sufficiently small for the forgetting tools to succeed.

By computing three different ontology modules and applying forgetting method to the 2016 core version of SNOMED CT, we created ontology abstractions for a sample of reference lists used in the health care sector in the UK (for the NHS Refsets). We used 165 concepts in the ERA Refset as class names for primary renal disorders in our evaluation. The results show that precomputing MEX-modules and minimal subsumption modules significantly reduce the size of extracts and significantly improve the performance of our forgetting tool. The returned uniform interpolants only consist of class and property names in the interpolation signatures. Because forgetting preserves semantic equivalence high precision can be achieved.

References


