Efficient Axiom Pinpointing with EL2MCS

M. Fareed Arif[1], Carlos Mencía[1], and Joao Marques-Silva[1,2]

[1] CASL, University College Dublin, Ireland
farif@ucdconnect.ie, carlos.mencia@ucd.ie, jpms@ucd.ie
[2] INESC-ID, IST, ULisboa, Portugal

Abstract. Axiom pinpointing consists in computing a set-wise minimal set of axioms that explains the reason for a subsumption relation in an ontology. Recently, an encoding of the classification of an $\mathcal{EL}^+$ ontology to a polynomial-size Horn propositional formula has been devised. This enables the development of a method for axiom pinpointing based on the analysis of unsatisfiable propositional formulas. Building on this earlier work, we propose a computation method, termed EL2MCS, that exploits an important relationship between minimal axiom sets and minimal unsatisfiable subformulas in the propositional domain. Experimental evaluation shows that EL2MCS achieves substantial performance gains over existing axiom pinpointing approaches for lightweight description logics.

1 Introduction

Axiom pinpointing consists in identifying a minimal set of axioms (MinA) that explains a given subsumption relation in an ontology. This problem is useful for debugging ontologies, and finds several application domains, including medical informatics [32,15,20]. Earlier axiom pinpointing algorithms [5,6] in lightweight Description Logics (i.e., $\mathcal{EL}$ and $\mathcal{EL}^+$) generate a (worst-case exponential-size) propositional formula and compute the MinAs by finding its minimal models, which is an NP-hard problem. More recently, a polynomial-size encoding is devised in [33,34] that encodes the classification of an $\mathcal{EL}^+$ ontology into a Horn propositional formula (i.e. it can be exponentially more compact than earlier work [5,6]). This encoding is exploited by the axiom pinpointing algorithm EL$^+$SAT [33,34], based on SAT methods [25] and SMT-like techniques [17]. Although effective at computing MinAs, these dedicated algorithms often fail to enumerate all MinAs to completion, or to prove that no additional MinA exists.

Building on this previous work, we present a new approach for axiom pinpointing in $\mathcal{EL}^+$ DLs, termed EL2MCS. It is based on a relationship between MinAs and minimal unsatisfiable subformulas (MUSes) of the Horn formula encoding [33,34]. The relationship between MUSes and MinAs makes it possible to benefit from the large recent body of work on extracting MUSes [14,8,9,24,29,16], but also minimal correction subsets (MCSes), as well as their minimal hitting set relationship [31,7,18], which for the propositional case allows for exploiting the performance of modern SAT solvers. The relationship between axiom pinpointing and MUS enumeration was also studied elsewhere independently [22], where the proposed approach iteratively computes prime implicants [12,24] instead of exploiting hitting set dualization.

Experimental results, considering instances from medical ontologies, show that EL2MCS significantly outperforms existing approaches [4,19,34,22].

The remainder of the paper is structured as follows. Section 2 introduces basic definitions and notation. Section 3 describes the propositional Horn encoding and our proposed axiom pinpointing approach. The experimental results are reported in Section 4 and Section 5 concludes the paper.

2 Preliminaries

2.1 Lightweight Description Logics

The standard definitions of $\mathcal{EL}^+$ are assumed [6,33]. Starting from a set $\mathsf{N_C}$ of *concept names* and a set $\mathsf{N_R}$ of *role names*, *concept descriptions* are defined inductively. A TBox is a finite set of *general concept inclusion* (GCI) and *role inclusion* (RI) axioms. For a TBox $\mathcal{T}$, $\mathsf{PC}_{\mathcal{T}}$ denotes the set of *primitive concepts* of $\mathcal{T}$, representing the smallest set of concepts that contains the top concept $\top$, and all the concept names in $\mathcal{T}$. $\mathsf{PR}_{\mathcal{T}}$ denotes the set of *primitive roles* of $\mathcal{T}$, representing all role names in $\mathcal{T}$. We assume standard set theoretic semantics [3] and the main inference problem for $\mathcal{EL}^+$ is concept subsumption [3,6]:

Definition 1 (Concept Subsumption). *Let C, D represent two $\mathcal{EL}^+$ concept descriptions and let $\mathcal{T}$ represent an $\mathcal{EL}^+$ TBox. C is subsumed by D w.r.t. $\mathcal{T}$ (denoted $C \sqsubseteq_{\mathcal{T}} D$) iff $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ in every model $\mathcal{I}$ of $\mathcal{T}$.*

Finding an explanation, termed *axiom pinpointing*, consists of computing a minimal axiom subset *(MinA)* that explains the subsumption relation. The standard definition of *MinA* [33] is assumed:

Definition 2 (MinA). *Let $\mathcal{T}$ be an $\mathcal{EL}^+$ TBox, and let $C, D \in \mathsf{PC}_{\mathcal{T}}$ be primitive concept names, with $C \sqsubseteq_{\mathcal{T}} D$. Let $\mathcal{S}$ be a subset of $\mathcal{T}$ be such that $C \sqsubseteq_{\mathcal{S}} D$. If $\mathcal{S}$ is such that $C \sqsubseteq_{\mathcal{S}} D$ and $C \not\sqsubseteq_{\mathcal{S}'} D$ for $\mathcal{S}' \subset \mathcal{S}$, then $\mathcal{S}$ is a* minimal axiom set *(MinA) w.r.t. $C \sqsubseteq_{\mathcal{T}} D$.*

2.2 Propositional Satisfiability

Standard propositional satisfiability (SAT) definitions are assumed [10]. Formulas are represented by $\mathcal{F}$, $\mathcal{M}$, $\mathcal{M}'$, $\mathcal{C}$ and $\mathcal{C}'$, but also by φ, ϕ and ψ. Horn formulas are such that every clause contains at most one positive literal. In this paper, we explore both MUSes and MCSes of propositional formulas.

Definition 3 (MUS). *$\mathcal{M} \subseteq \mathcal{F}$ is a* Minimal Unsatisfiable Subformula *(MUS) of $\mathcal{F}$ iff $\mathcal{M}$ is unsatisfiable and $\forall_{\mathcal{M}' \subsetneq \mathcal{M}}\ \mathcal{M}'$ is satisfiable.*

Definition 4 (MCS). *$\mathcal{C} \subseteq \mathcal{F}$ is a* Minimal Correction Subformula *(MCS) of $\mathcal{F}$ iff $\mathcal{F} \setminus \mathcal{C}$ is satisfiable and $\forall_{\mathcal{C}' \subsetneq \mathcal{C}}\ \mathcal{F} \setminus \mathcal{C}'$ is unsatisfiable.*

A well-known result, which will be used in the paper is the minimal hitting set relationship between MUSes and MCSes of an unsatisfiable formula $\mathcal{F}$ [31,11,7,18].

Theorem 1. *Let $\mathcal{F}$ be unsatisfiable. Then, each MCS of $\mathcal{F}$ is a minimal hitting set of the MUSes of $\mathcal{F}$ and each MUS of $\mathcal{F}$ is a minimal hitting set of the MCSes of $\mathcal{F}$.*

A *partial MaxSAT*, formula φ is that partitioned into a set of hard (φ_H) and soft (φ_S) clauses, i.e. $\varphi = \{\varphi_H, \varphi_S\}$. Hard clauses must be satisfied while soft clauses can be relaxed. We have used partial MaxSAT encoding and enumeration of MUSes [7,18] using minimal hitting set duals [31,11,7,18] in our proposed solution.

3 Computation Technique and Tool Overview (EL2MCS)

This section introduces the main organization of our approach. It works over the propositional Horn encoding used in EL$^+$SAT [33,34], and exploits a close relationship between MinAs and MUSes.

3.1 Horn Formula Encoding

In EL$^+$SAT, the Horn formula $\phi^{\text{all}}_{\mathcal{T}(\text{po})}$ mimics the classification of TBox $\mathcal{T}$ and is constructed as follows [33,34]:

1. For every axiom, create an axiom selector propositional variable $s_{[a_i]}$. For trivial GCI of the form $C \sqsubseteq C$ or $C \sqsubseteq \top$, $s_{[a_i]}$ is constant `true`.
2. During the execution of the classification algorithm [3,6], for every application of a rule (concretely r) generating some assertion (concretely a_i), add to $\phi^{\text{all}}_{\mathcal{T}(\text{po})}$ a clause of the form,
$$\left(\bigwedge_{a_j \in \mathsf{ant}(r)} s_{[a_j]} \right) \rightarrow s_{[a_i]}$$
where $s_{[a_i]}$ is the selector variable for a_i and $\mathsf{ant}(r)$ are the antecedents of a_i with respect to a completion rule r.

For axiom pinpointing the SAT-based algorithms [33,34] , exploiting the ideas from early work on SAT solving [25] and AllSMT [17], compute MinAs for any subsumption relation (i.e., $C_i \sqsubseteq D_i$) using the list of assumption variables $\{\neg s_{[C_i \sqsubseteq D_i]}\} \cup \{s_{[ax_i]} \,|\, ax_i \in \mathcal{T}\}$. The following theorem is fundamental for this work [33,34], and is extended in the next section to relate MinAs with MUSes of propositional formulas.

Theorem 2 (Theorem 3 in [34]). *Given an $\mathcal{EL}^+$ TBox $\mathcal{T}$, for every $\mathcal{S} \subseteq \mathcal{T}$ and for every pair of concept names $C, D \in \mathsf{PC}_{\mathcal{T}}$, $C \sqsubseteq_{\mathcal{S}} D$ if and only if the Horn propositional formula $\phi^{\text{all}}_{\mathcal{T}(\text{po})} \wedge (\neg s_{[C \sqsubseteq D]}) \wedge_{ax_i \in \mathcal{S}} (s_{[ax_i]})$ is unsatisfiable.*

3.2 MinAs as MUSes

Although not explicitly stated, the relation between axiom pinpointing and MUS extraction has been apparent in earlier work [6,33,34].

Theorem 3 ([1]). *Given an $\mathcal{EL}^+$ TBox $\mathcal{T}$, for every $\mathcal{S} \subseteq \mathcal{T}$ and for every pair of concept names $C, D \in \mathsf{PC}_{\mathcal{T}}$, $\mathcal{S}$ is a **MinA** of $C \sqsubseteq_{\mathcal{S}} D$ if and only if the Horn propositional formula $\phi^{\text{all}}_{\mathcal{T}(\text{po})} \wedge (\neg s_{[C \sqsubseteq D]}) \wedge_{ax_i \in \mathcal{S}} (s_{[ax_i]})$ is **minimally unsatisfiable**.*

Based on Theorem 3 and the MUS enumeration approach in [18], we can now outline our axiom pinpointing approach.

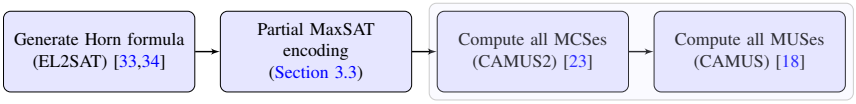

Fig. 1: The EL2MCS tool

3.3 Axiom Pinpointing using MaxSAT

As described earlier, the axiom pinpointing algorithm [33,34] explicitly enumerates the selection variables (i.e., $s_{[ax_i]}$) in a AllSMT-inspired approach [17]. In contrast, our approach is to model the problem as partial maximum satisfiability (MaxSAT), and enumerate over the MUSes of the MaxSAT problem formulation. Therefore, all clauses in $\phi^{\text{all}}_{\mathcal{T}(\text{po})}$ are declared as hard clauses, Observe that, by construction, $\phi^{\text{all}}_{\mathcal{T}(\text{po})}$ is satisfiable. In addition, the constraint $C \sqsubseteq_{\mathcal{T}} D$ is encoded with another hard clause, namely $(\neg s_{[C \sqsubseteq_{\mathcal{T}} D]})$. Finally, the variable $s_{[ax_i]}$ associated with each axiom ax_i denotes a *unit soft clause*. The intuitive justification is that the goal is to include as many axioms as possible, leaving out a minimal set which, if included, would cause the complete formula to be unsatisfiable. Thus, each of these sets represents an MCS of the MaxSAT problem formulation, but also a minimal set of axioms that needs to be dropped for the subsumption relation not to hold (i.e. a *diagnosis* [20]). MCS enumeration can be implemented with a MaxSAT solver [18,27] or with a dedicated algorithm [23]. It is well-known (e.g. see Theorem 1) that MCSes are minimal hitting sets of MUSes, and MUSes are minimal hitting sets of MCSes [31,11,7,18]. Thus, we use explicit minimal hitting set duality to obtain the MUSes we are looking for, starting from the previously computed MCSes.

3.4 EL2MCS Tool

The organization of the EL2MCS tool is shown in Figure 1. The first step is similar to EL$^+$SAT [33,34] in that a propositional Horn formula is generated. The next step, however, exploits the ideas in Section 3.3, and generates a partial MaxSAT encoding. We can now enumerate the MCSes of the partial MaxSAT formula using the CAMUS2 tool [23][1]. The final step is to exploit minimal hitting set duality for computing all the MUSes given the set of MCSes [18]. This is achieved with the CAMUS tool[2]. The hypergraph traversal computation tools, shd [28] and MTminer [13], could be used instead in this phase. It should be observed that, although MCS enumeration uses CAMUS2 (a modern implementation of the MCS enumerator in CAMUS [18], capable of handling partial MaxSAT formulae), alternative MCS enumeration approaches were considered [23] but found not to be as efficient.

[1] Available from `http://logos.ucd.ie/web/doku.php?id=mcsls`.
[2] Available from `http://sun.iwu.edu/~mliffito/camus/`.

	vs EL$^+$SAT	vs SATPin	vs CEL	vs JUST
#Wins / #Losses	359 / 106	353 / 114	379 / 18	236 / 28
%Wins / %Losses	71.8% / 21.2%	70.6% / 22.8%	96.2% / 4.5%	80.8% / 9.6%

Table 1: Summary of results (wins/losses). Draws are not represented.

4 Experiment Evaluation

This section presents an empirical evaluation of EL2MCS[3], which is compared to the state-of-the-art tools CEL [4], JUST [19], EL$^+$SAT [34] and SATPin [22]. EL$^+$SAT and SATPin are SAT-based approaches, whereas CEL and JUST use dedicated reasoners.

The medical ontologies used in the experiments are GALEN [30] (two variants: FULL-GALEN and NOT-GALEN), Gene [2], NCI [35] and SNOMED-CT [36]. As in earlier work [34], for each ontology 100 subsumption query instances were considered. So, there are 500 instances. In addition, for the SAT-based tools, including EL2MCS, the instances were simplified with the *cone-of-influence* (COI) reduction technique[4] [34]. CEL and JUST use their own similar simplification techniques. The comparison with CEL and JUST imposes additional constraints. CEL reports at most 10 MinAs, so only 397 instances with up to 10 MinAs were considered in the comparison with CEL. JUST is only able to handle a subset of $\mathcal{EL}^+$, so the comparison with JUST only considers 292 instances it can return correct results. The experiments were performed on a Linux Cluster (2GHz), with a time limit of 3600s.

By the time limit, out of the 500 instances, EL$^+$SAT solves 241, SATPin solves 458 and EL2MCS solves 470. For the few instances EL2MCS does not solve, it computes thousands of MCSes by the time limit without reporting any MinA. In these cases, EL$^+$SAT and SATPin are able to return some MinAs, although not achieving complete enumeration. Regarding the comparison with CEL, out of 397 instances, CEL solves 394 and EL2MCS solves all of them. Compared with JUST, out of the 292 instances considered, JUST solves 242 and EL2MCS solves 264. It is worth mentioning that there is no instance some tool is able to solve and EL2MCS is not.

Table 1 compares EL2MCS with the other tools in terms of the number of instances it performed better and worse (wins/losses). Unsolved instances where some method computed some MinAs and EL2MCS did not, are counted as losses. As we can observe, in a majority of the cases, EL2MCS performs better than any other tool.

Figure 2 shows four scatter plots with a pairwise comparison of EL2MCS and each other tool in terms of their running times. They reveal very significant differences in favor of EL2MCS in all cases. EL2MCS is remarkably faster than any other tool for most instances, in many cases with performance gaps of more than one order of magnitude. The greatest advantages are over EL$^+$SAT, CEL and JUST. SATPin performs better than other alternatives, but still EL2MCS outperforms it consistently as well.

Summing up, not only EL2MCS is able to solve more instances than the state-of-the-art tools, but also it is much faster than any other alternative for most instances.

[3] Available from `http://logos.ucd.ie/web/doku.php?id=el2mcs`.

[4] We observed similar differences in performance between EL2MCS and the other tools when either not simplifying the instances or considering other simplification techniques [34].

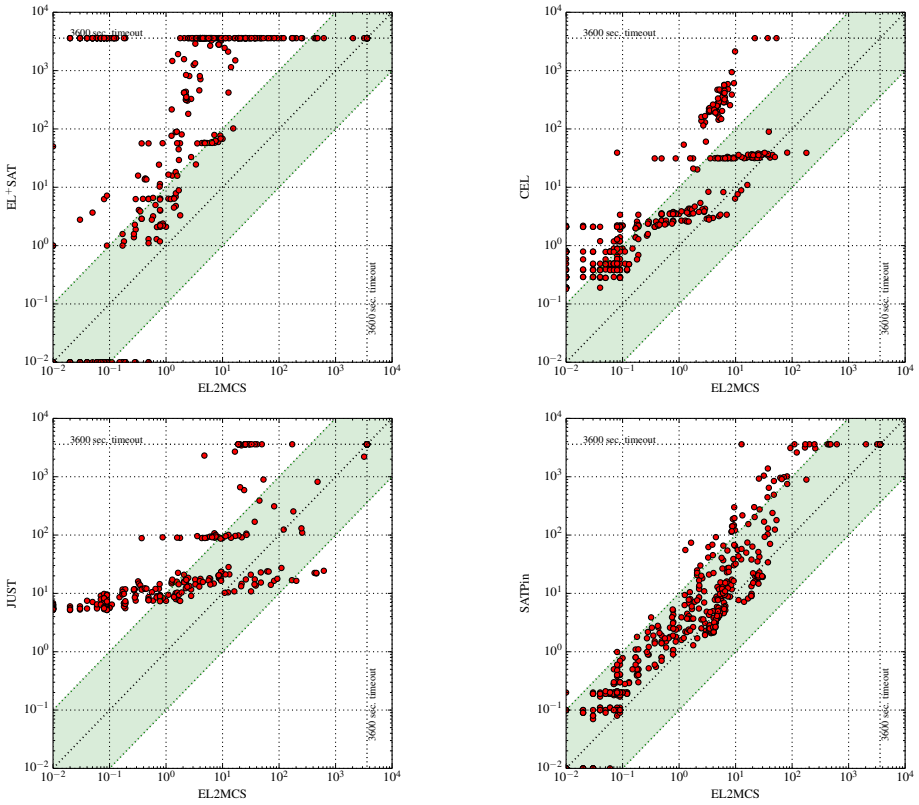

Fig. 2: Scatter plots comparing EL2MCS with EL$^+$SAT, CEL, JUST and SATPin.

5 Conclusions & Future Work

This paper presents the EL2MCS tool for axiom pinpointing of $\mathcal{EL}^+$ ontologies. Building on previous work [33,34], EL2MCS exploits a close relationship between MinAs and MUSes of propositional formulas, and instruments an efficient algorithm that relies on explicit minimal hitting set dualization of MCSes and MUSes of unsatisfiable formulas. Experimental results over well-known benchmarks from medical ontologies, reveal that EL2MCS signficantly outperforms the state of the art, thus constituting a very effective alternative for this problem. A natural research direction is to attempt to improve EL2MCS by substituting some of its parts by other advanced novel alternatives (e.g. MCS extraction and enumeration [21,26]).

Acknowledgement. We thank R. Sebastiani and M. Vescovi, for authorizing the use of EL$^+$SAT [34]. We thank N. Manthey and R. Peñaloza, for bringing SATPin [22] to our attention, and for allowing us to use their tool. This work is partially supported by SFI PI grant BEACON (09/IN.1/I2618), by FCT grant POLARIS (PTDC/EIA-CCO/123051/-2010), and by national funds through FCT with reference UID/CEC/50021/2013.

References

1. M. Fareed Arif and Joao Marques-Silva. Towards efficient axiom pinpointing of EL+ ontologies. *CoRR*, abs/1503.08454, 2015. Available from http://arxiv.org/abs/1503.08454.
2. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, and et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
3. Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the $\mathcal{EL}$ envelope. In *IJCAI*, pages 364–369, 2005.
4. Franz Baader, Carsten Lutz, and Boontawee Suntisrivaraporn. CEL - A polynomial-time reasoner for life science ontologies. In *IJCAR*, pages 287–291, 2006.
5. Franz Baader, Rafael Peñaloza, and Boontawee Suntisrivaraporn. Pinpointing in the description logic $\mathcal{EL}$. In *DL*, 2007.
6. Franz Baader, Rafael Peñaloza, and Boontawee Suntisrivaraporn. Pinpointing in the description logic $\mathcal{EL}^+$. In *KI*, pages 52–67, 2007.
7. James Bailey and Peter J. Stuckey. Discovery of minimal unsatisfiable subsets of constraints using hitting set dualization. In *PADL*, pages 174–186, 2005.
8. Anton Belov, Marijn Heule, and Joao Marques-Silva. MUS extraction using clausal proofs. In *SAT*, pages 48–57, 2014.
9. Anton Belov, Inês Lynce, and Joao Marques-Silva. Towards efficient MUS extraction. *AI Commun.*, 25(2):97–116, 2012.
10. Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors. *Handbook of Satisfiability*, volume 185, 2009.
11. Elazar Birnbaum and Eliezer L. Lozinskii. Consistent subsets of inconsistent systems: structure and behaviour. *J. Exp. Theor. Artif. Intell.*, 15(1):25–46, 2003.
12. Aaron R. Bradley and Zohar Manna. Checking safety by inductive generalization of counterexamples to induction. In *FMCAD*, pages 173–180, 2007.
13. Céline Hébert, Alain Bretto, and Bruno Crémilleux. A data mining formalization to improve hypergraph minimal transversal computation. *Fundammenta Informaticae*, 80(4):415–433, 2007.
14. Ulrich Junker. QuickXplain: Preferred explanations and relaxations for over-constrained problems. In *AAAI*, pages 167–172, 2004.
15. Aditya Kalyanpur, Bijan Parsia, Evren Sirin, and Bernardo Cuenca Grau. Repairing unsatisfiable concepts in owl ontologies. pages 170–184, 2006.
16. Jean-Marie Lagniez and Armin Biere. Factoring out assumptions to speed up MUS extraction. In *SAT*, pages 276–292, 2013.
17. Shuvendu K. Lahiri, Robert Nieuwenhuis, and Albert Oliveras. SMT techniques for fast predicate abstraction. In *CAV*, pages 424–437, 2006.
18. Mark H. Liffiton and Karem A. Sakallah. Algorithms for computing minimal unsatisfiable subsets of constraints. *J. Autom. Reasoning*, 40(1):1–33, 2008.
19. Michel Ludwig. Just: a tool for computing justifications w.r.t. el ontologies. In *ORE 2014*, volume 1207, pages 1–7, 2014.
20. Michel Ludwig and Rafael Peñaloza. Error-tolerant reasoning in the description logic $\mathcal{EL}$. In *JELIA*, pages 107–121, 2014.
21. Yuri Malitsky, Barry O'Sullivan, Alessandro Previti, and Joao Marques-Silva. Timeout-sensitive portfolio approach to enumerating minimal correction subsets for satisfiability problems. In *ECAI*, pages 1065–1066, 2014.
22. Norbert Manthey and Rafael Peñaloza. Exploiting SAT technology for axiom pinpointing. Technical Report LTCS 15-05, Chair of Automata Theory, Institute of Theoretical Computer

Science, Technische Universität Dresden, April 2015. Available from https://ddll.inf.tu-dresden.de/web/Techreport3010.
23. Joao Marques-Silva, Federico Heras, Mikoláš Janota, Alessandro Previti, and Anton Belov. On computing minimal correction subsets. In *IJCAI*, 2013.
24. Joao Marques-Silva, Mikoláš Janota, and Anton Belov. Minimal sets over monotone predicates in Boolean formulae. In *CAV*, pages 592–607, 2013.
25. Joao Marques-Silva, Inês Lynce, and Sharad Malik. Conflict-driven clause learning SAT solvers. In Biere et al. [10], pages 131–153.
26. Carlos Mencía, Alessandro Previti, and Joao Marques-Silva. Literal-based MCS extraction. In *IJCAI*, 2015.
27. António Morgado, Mark H. Liffiton, and Joao Marques-Silva. MaxSAT-based MCS enumeration. In *HVC*, pages 86–101, 2012.
28. Keisuke Murakami and Takeaki Uno. Efficient algorithms for dualizing large-scale hypergraphs. *CoRR*, abs/1102.3813, 2011.
29. Alexander Nadel, Vadim Ryvchin, and Ofer Strichman. Efficient MUS extraction with resolution. In *FMCAD*, pages 197–200, 2013.
30. Alan L. Rector and Ian R. Horrocks. Experience building a large, re-usable medical ontology using a description logic with transitivity and concept inclusions. In *Workshop on Ontological Engineering*, pages 414–418, 1997.
31. Raymond Reiter. A theory of diagnosis from first principles. *Artif. Intell.*, 32(1):57–95, 1987.
32. Stefan Schlobach, Zhisheng Huang, Ronald Cornet, and Frank van Harmelen. Debugging incoherent terminologies. *J. Autom. Reasoning*, 39(3):317–349, 2007.
33. Roberto Sebastiani and Michele Vescovi. Axiom pinpointing in lightweight description logics via Horn-SAT encoding and conflict analysis. In *CADE*, pages 84–99, 2009.
34. Roberto Sebastiani and Michele Vescovi. Axiom pinpointing in large $\mathcal{EL}^+$ ontologies via SAT and SMT techniques. Technical Report DISI-15-010, DISI, University of Trento, Italy, April 2015. Under Journal Submission. Available as http://disi.unitn.it/~rseba/elsat/elsat_techrep.pdf.
35. Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.
36. Kent A. Spackman, Keith E. Campbell, and Roger A. Côté. SNOMED RT: a reference terminology for health care. In *AMIA*, 1997.