The Empirical Robustness of Description Logic Classification

Rafael S. Gonçalves, Nicolas Matentzoglu, Bijan Parsia, and Uli Sattler

School of Computer Science, University of Manchester, Manchester, United Kingdom

Abstract. In spite of the recent renaissance in lightweight description logics (DLs), many prominent DLs, such as that underlying the Web Ontology Language (OWL), have high worst case complexity for their key inference services. Modern reasoners have a large array of optimization, tuned calculi, and implementation tricks that allow them to perform very well in a variety of application scenarios, even though the complexity results ensure that they will perform poorly for some inputs. For users, the key question is how often they will encounter those pathological inputs in practice, that is, how robust are reasoners. We attempt to determine this question for classification of existing ontologies as they are found on the Web. It is a fairly common user task to examine ontologies published on the Web as part of their development process. Thus, the robustness of reasoners in this scenario is both directly interesting and provides some hints toward answering the broader question. From our experiments, we show that the current crop of OWL reasoners, in collaboration, is very robust against the Web.

1 Motivation

A serious concern about both versions 1 [4] and 2 [3] of the Web Ontology Language (OWL) is that the underlying description logics ($\mathcal{SHOIQ}$ and $\mathcal{SROIQ}$) exhibit extremely bad worst case complexity (NEXPTIME and 2NEXPTIME) for their key inference services. While since the mid-1990s, highly optimized description logic reasoners have been exhibiting rather good performance in real cases, even in those more constrained cases there are ontologies (such as Galen) which have proved impossible to process for over a decade. Indeed, concern with such pathology stimulated a renaissance of research into tractable description logics with the $\mathcal{EL}$ family [1] and the DL Lite [2] family being incorporated as special "profiles" of OWL 2. However, even though the number of ontologies available on the Web has grown enormously since the standardization of OWL, it is still unclear how robust modern, highly optimized reasoners are to such input. Anecdotal evidence suggests that pathological cases are common enough to cause problems, however, systematic evidence has been scarce.

In this paper we investigate the question of whether modern, highly-optimized description logic reasoners are *robust* over Web input. The general intuition of a robust system is that it is *resistant to failure in the face of a range of input*. For any particular robustness determination, one must decide: *1)* the range of input, *2)* the functional or non-functional properties of interest, and *3)* what counts as failure. The instantiation of these parameters strongly influences robustness judgements, with the very same reasoner being highly robust under one scenario and very non-robust under another. For our

current purposes, the key scenario is that an ontology engineer, using a tool like Protégé [6], is inspecting ontologies published on the Web with an eye to possible reuse, and, as is common, they wish to classify the ontology using a standard OWL 2 DL reasoner as part of their evaluation. This scenario yields the following constraints: *1)* for input, we examine Web-based corpora, *2)* functional: acceptance (will the reasoner load and process the ontology); non-functional: performance (i.e., will the reasoner complete classification before the ontology engineer gives up), *3)* w.r.t. acceptance, failure means either rejecting the input or crashing while processing, and we might reasonably expect an engineer to wait up to 2 hours if the ontology seems "worth it". If a reasoner (or a set of reasoners) is successful for 90% of a corpus, we count that reasoner as robust over that corpus, with 95% and 99% indicating "strong" and "extreme" robustness. While these levels are clearly arbitrary (as is the timeout), they provide a framework to set expectations. Robustness under these assumptions does not ensure robustness under other assumptions (e.g., over subsets of these ontologies as experienced during development or over a more stringent time constraint), yet they are challenging enough that it was unclear to us *ex ante* whether any reasoner would be robust for any corpus.

In fact, we find that the reasoners are robust or near robust for most of the cases we examine, including for lower timeouts. More significantly, if we take the best result for each ontology (which represents a kind of "meta-reasoner", where our test reasoners are run in parallel), then the *set* of reasoners is extremely robust over all corpora. Thus, in a fairly precise, if limited, sense, we demonstrate that classification over OWL ontologies (even those based on highly expressive description logics, such as $\mathcal{SHOIQ}$ and $\mathcal{SROIQ}$) is practical, even despite the worst case being intractable in some cases.

2 Results

Overall we have processed a total of 1,071 ontologies, the largest such reasoner benchmark (similar benchmarks typically use at most a few hundred ontologies, e.g., the recent study in [5]), having found that amongst the 4 tested reasoners Pellet is the most robust of all (see Table 1). Surprisingly, Pellet is followed by JFact on our robustness test, due to having far less errors than FaCT++. HermiT and FaCT++ have the same overall robustness, but FaCT++ has less errors and higher impatient robustness.

While Pellet is the most robust reasoner, we urge some caution in that reading. In particular, this does not mean that Pellet will always do best or even perform reasonably. In fact, it may timeout where other reasoners finish reasonably fast. The set of reasoners (taken together and considering the best results) is extremely robust across the board (for each reasoner's contribution to the best case reasoner, see Figure 1). Thus, we have strong empirical evidence that the *ontologies* on the Web do not supply any *in principle* intractable cases, but only cases which are difficult for particular reasoners.

Note that FaCT++ and JFact fail to process several ontologies due to poor support for OWL 2 datatypes. Both of these reasoners, as well as HermiT, seem to have little support for OWL 1 datatypes. By removing the non OWL 2 datatype errors, we would end up with FaCT++ being the most robust w.r.t. OWL 2, followed by HermiT and Pellet. That is, if we restrict the test corpus to those ontologies that use only datatypes from the OWL 2 datatype map, then FaCT++ would be the most robust reasoner.

	Pellet	HermiT	JFact	FaCT++	Best Combo	Worst Combo
Very Easy	787 (73%)	706 (66%)	741 (69%)	784 (73%)	878 (82%)	645 (60.2%)
Easy	116 (11%)	112 (10%)	103 (10%)	55 (5%)	73 (7%)	102 (9.5%)
Medium	83 (8%)	65 (6%)	43 (4%)	94 (9%)	101 (9%)	45 (4.2%)
Hard	24 (2%)	53 (5%)	76 (7%)	7 (1%)	6 (1%)	75 (7.0%)
Very Hard	6 (1%)	4 (0%)	1 (0%)	4 (0%)	4 (0%)	3 (0.3%)
Timeout	29 (3%)	14 (1%)	16 (1%)	15 (1%)	9 (1%)	25 (2.3%)
Errors	26 (2%)	117 (11%)	91 (8%)	112 (10%)	0 (0%)	176 (16.4%)
Total (excl. Errors)	1016	940	964	944	1062	870
Total (incl. Errors)	1071	1071	1071	1071	1071	1071
Impatient Robustness	92%	82% [90%]	83%	87% [96%]	98%	74% [87%]
Overall Robustness	95%	88% [96%]	90%	88% [97%]	99%	81% [96%]

Table 1: Binning of all three corpora: BioPortal, NCIt (2013), and Web crawl. Under robustness rows, values in square brackets indicate robustness w.r.t. OWL 2 alone.

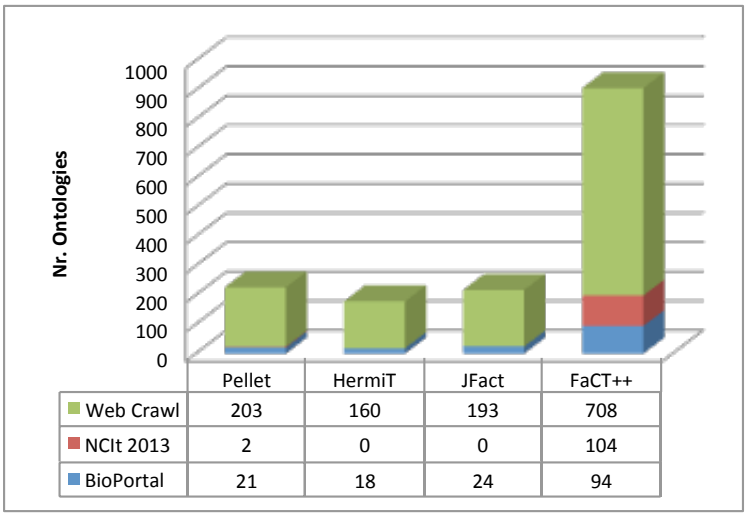

Fig. 1: Number of times each reasoner outperforms all other reasoners in each corpus.

From Figure 1 we see that FaCT++ outperforms other reasoners on many occasions, but, due to the high number of errors thrown, its robustness w.r.t. our input data is not as good as this figure might indicate. In Figure 2 we show the frequency with which reasoners are the worst case in each corpus: Notice that FaCT++ is, overall, less often the worst reasoner, followed by HermiT. However, HermiT and JFact both dominate the worst cases in the NCIt corpus. Pellet, while being most often the worst case reasoner in the Web Crawl corpus, is so (in many cases) by a mere fraction of a second; as pointed out it is the most robust for that corpus.

It is clear that deriving a sensible ranking even simply using average or total time is not straightforward. Our results have rather strong implications for reasoner experi-

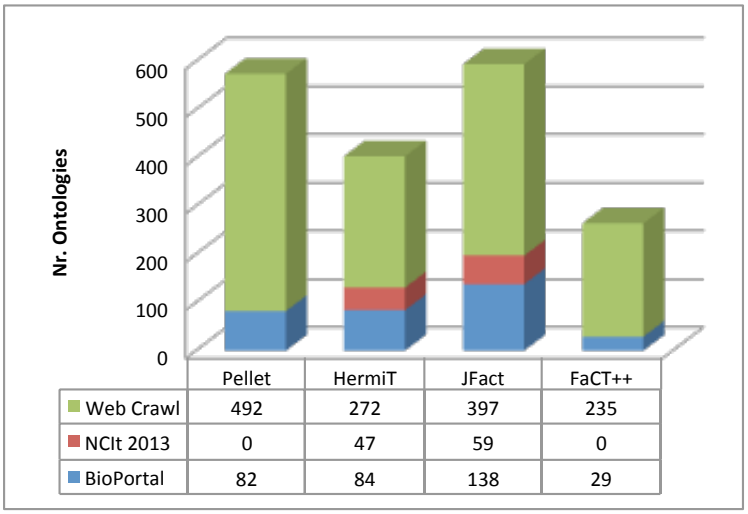

Fig. 2: Number of times that each reasoner equals the worst case, for each corpus.

ments, especially those purporting to show the advantages of an optimisation or a technique or an implementation: The space is very complex and it is very easy to simultaneously generate a biased sample for one system and against another. Even simple, seemingly innocuous things like timeouts and classification failures require tremendous care in handling. If results are going to be meaningful across papers we need to converge on experimental inputs, methods, and reporting forms.

References

1. Baader, F., Brandt, S., Lutz, C.: Pushing the EL envelope. In: Proc. of IJCAI-05 (2005)
2. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Tractable reasoning and efficient query answering in description logics: The DL-Lite family. J. of Automated Reasoning 39(3), 385–429 (2007)
3. Cuenca Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P.F., Sattler, U.: OWL 2: The next step for OWL. J. of Web Semantics (2008)
4. Horrocks, I., Patel-Schneider, P.F., van Harmelen, F.: From $\mathcal{SHIQ}$ and RDF to OWL: The making of a web ontology language. J. of Web Semantics 1(1), 7–26 (2003)
5. Kang, Y.B., Li, Y.F., Krishnaswamy, S.: Predicting reasoning performance using ontology metrics. In: Proc. of ISWC-12 (2012)
6. Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A.: The Protégé OWL plugin: An open development environment for semantic web applications. In: Proc. of ISWC-04 (2004)