

SAT-based PAC Learning of Description Logic Concepts (Extended Abstract)

Balder ten Cate¹, Maurice Funk^{2,3}, Jean Christoph Jung⁴ and Carsten Lutz^{2,3}

¹ILLC, University of Amsterdam

²Leipzig University

³Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI)

⁴TU Dortmund University

Abstract

We propose *bounded fitting* as a scheme for learning description logic concepts in the presence of ontologies. Two main advantages are (1) theoretical guarantees regarding the generalization of the learned concepts to unseen examples in the sense of PAC learning and (2) the fact that implementation can leverage SAT solvers in a natural way. We also present our system SPELL which implements bounded fitting based on a SAT solver and compare its performance to a state-of-the-art learner.


Keywords

PAC learning, Description Logic, Concept Learning

The manual curation of knowledge bases (KBs) is time consuming and expensive, making learning-based approaches to knowledge acquisition an attractive alternative. In description logics (DLs), *concepts* are a fundamental class of expressions that are used as a central building block for ontologies and also as queries to KBs. Consequently, the subject of learning DL concepts from labeled data examples has received great interest, resulting in various implemented systems such as DL-Learner, DL-Foil, and YINYANG [1, 2, 3]. These systems take as input a set of positively and negatively labeled examples and an ontology \mathcal{O} , and try to construct a concept that fits the examples w.r.t. \mathcal{O} . This is related to the *fitting problem*, asking to decide the existence of a fitting concept, which has also been studied intensely [4, 5, 6].

In this extended abstract we report about the recent publication [7], see also [8] for technical details and proofs. We propose a new approach to concept learning in DLs that we call *bounded fitting*, inspired by both bounded model checking as known from systems verification [9] and by Occam algorithms from computational learning theory [10]. The idea of bounded fitting is to search for a fitting concept of bounded size, iteratively increasing the size bound until a fitting is found. This approach has two main advantages, which we discuss next.


First, it comes with formal guarantees regarding the generalization of the returned concept from the training data to previously unseen data. This is formalized by Valiant's framework of *probably approximately correct (PAC) learning* [11]. Given sufficiently many data examples sampled from an unknown distribution, bounded fitting returns a concept that with high probability δ has a classification error bounded by some small ϵ on examples drawn according

 DL 2023: 36th International Workshop on Description Logics, September 2–4, 2023, Rhodes, Greece

 b.d.tencate@uva.nl (B. ten Cate); maurice.funk@uni-leipzig.de (M. Funk); jean.jung@tu-dortmund.de (J. C. Jung); carsten.lutz@uni-leipzig.de (C. Lutz)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

to the same distribution. It is well-known that PAC learning is intimately linked to Occam algorithms which guarantee to find a hypothesis of small size [10, 12]. By design, algorithms following the bounded fitting paradigm are Occam, and as a consequence the number of examples needed for generalization depends only linearly on $1/\delta$, and the size of the target concept to be learned, and log-linearly on $1/\epsilon$. This generalization guarantee holds independently of the DL used to formulate concepts and ontologies.

The second advantage is that, in important cases, bounded fitting enables learning based on SAT solvers and thus leverages the practical efficiency of these systems. In particular, this is the case for concepts formulated in \mathcal{EL} and ontologies formulated in \mathcal{ELH}^r . In this case, the *size-restricted fitting problem*, which is defined like the fitting problem except that the maximum size of fitting concepts to be considered is given as an additional input (in unary), is NP-complete [13]. This situation indeed suggests to use a SAT solver. For comparison, we mention that the unbounded fitting problem is EXPTIME-complete in this case [5]. The use of a SAT-solver is further justified by the fact that there is no *polynomial time* algorithm for learning \mathcal{EL} -concepts with PAC guarantees, unless $\text{RP} = \text{NP}$ [14, 15].

As a further contribution, we analyze the generalization ability of other relevant approaches to constructing fitting \mathcal{EL} -concepts. We start with algorithms that return fittings that are ‘prominent’ from a logical perspective in that they are most specific or most general or of minimum quantifier depth among all fittings. Algorithms with such characteristics and their applications are discussed in [16]. Notably, constructing fittings via direct products of positive examples yields most specific fittings [17, 18]. Our result is that, even without ontologies, these types of algorithms are not *sample-efficient*, that is, no polynomial amount of positive and negative examples is sufficient to achieve generalization in the PAC sense.

We next turn to algorithms based on so-called downward refinement operators which underlie all implemented DL learning systems that we are aware of. We consider two natural such operators that are rather similar to one another and combine them with a breadth-first search strategy. The first operator can be described as exploring ‘most-general specializations’ of the current hypotheses and the second one does the same, but is made ‘artificially Occam’ (with, most likely, a negative impact on practicality). We prove that while the first operator does not lead to a sample-efficient algorithm (even without ontologies), the second one does. This leaves open whether or not implemented systems based on refinement operators admit generalization guarantees, as they implement complex heuristics and optimizations.

As our final contribution we present SPELL (for *SAT-based PAC \mathcal{EL} concept Learner*), a SAT-based system that implements bounded fitting of \mathcal{EL} -concepts under \mathcal{ELH}^r -ontologies.¹ We evaluate SPELL on several datasets and compare it to the only other available learning system for \mathcal{EL} that we are aware of, the *\mathcal{EL} tree learner (ELTL)* component of the *DL-Learner* system [1]. We find that the running time of SPELL is almost always significantly lower than that of ELTL. Since, as we also show, it is the size of the target concept that has most impact on the running time, this means that SPELL can learn larger target queries than ELTL. We also analyze the relative strengths and weaknesses of the two approaches, identifying classes of inputs on which one of the systems performs significantly better than the other one. Finally, we make initial experiments regarding generalization, where both systems generalize well to unseen data, even

¹Available at <https://github.com/spell-system/SPELL>.

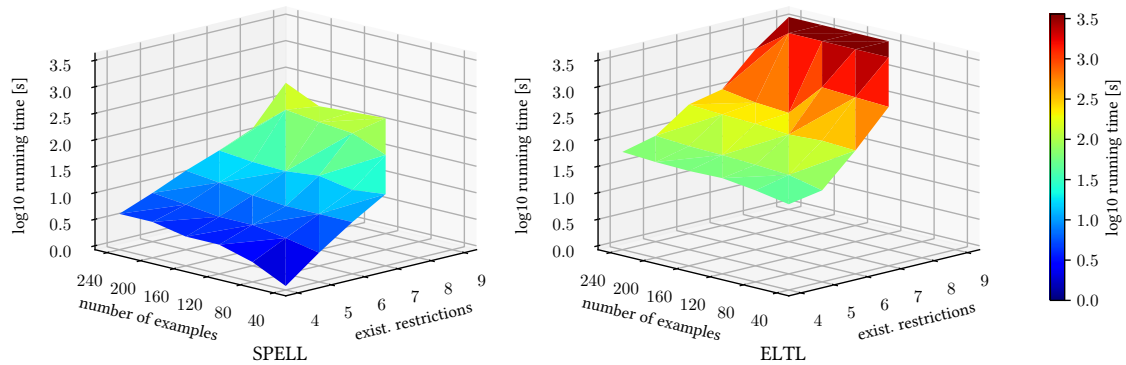


Figure 1: Yago experiment, dark red area indicates timeout (60min)

on very small samples. While this is expected for SPELL, for ELTL it may be due to the fact that some of the heuristics prefer fittings of small size, which might make ELTL an Occam algorithm.

Experimental Results

We refer to [7] for theoretical results and summarize here some of our experimental results. But first we make precise the bounded fitting approach to concept learning.

Definition 1. Let \mathcal{L} be an ontology language, \mathcal{C} a concept language and let \mathfrak{A} be an algorithm for the size-restricted fitting problem for \mathcal{C} -concepts under \mathcal{L} -ontologies. Then $\text{BOUNDED-FITTING}_{\mathfrak{A}}$ is the algorithm that, given a collection of labeled data examples E and an \mathcal{L} -ontology \mathcal{O} , runs \mathfrak{A} with input (E, \mathcal{O}, s) to decide whether there is a $C \in \mathcal{C}$ with $\|C\| \leq s$ that fits E w.r.t. \mathcal{O} , for $s = 1, 2, 3, \dots$, returning a fitting concept as soon as it finds one.

As mentioned in the introduction, the size-restricted fitting problem for \mathcal{EL} -concepts under \mathcal{ELH}^r -ontologies is NP-complete. On top of that, it admits a natural translation to SAT. Along with some optimizations, e.g., to break symmetries in the solution space for the resulting SAT formulas, we have implemented this translation in the system SPELL. We have evaluated SPELL on several newly crafted benchmarks and compared it to the ELTL component of the DL-Learner system [1]. Existing benchmarks did not suit our purpose as they aim at learning concepts that are formulated in more expressive DLs of the \mathcal{ALC} family; as a consequence, a fitting \mathcal{EL} concept almost never exists.

The first benchmark uses the Yago 4 knowledge base [19]. The smallest version of Yago 4 is still huge and contains over 40 million assertions. We extracted a fragment of 12 million assertions that focusses on movies and famous persons. We then systematically vary the number of labeled examples and the size of the target \mathcal{EL} -concepts. The latter take the form $C_n = \exists \text{actor} . \bigwedge_{i=1}^n r_i . \top$ where each r_i is a role name that represents a property of actors in Yago and n is increased to obtain larger queries. The positive examples are selected by querying Yago with C_n and the negative examples by querying Yago with generalizations of C_n . The results are presented in Figure 1. They show that the size of the target concept has a strong

Table 1

OWL2Bench running times [s], TO: >60min

	o2b-1	o2b-2	o2b-3	o2b-4	o2b-5	o2b-6
ELTL	TO	TO	274	580	28	152
SPELL	< 1	< 1	< 1	< 1	< 1	< 1

Table 2

Generalization experiment accuracies

Sample Size	5	10	15	20	25	30	35	40	45	50	55	60	65
ELTL	0.77	0.78	0.85	0.85	0.86	0.89	0.90	0.96	0.96	0.96	0.96	0.98	0.98
SPELL	0.80	0.81	0.84	0.85	0.86	0.86	0.89	0.97	0.98	0.98	0.98	0.98	0.98

impact on the running time whereas the impact of the number of positive and negative examples is much more modest. We also find that SPELL performs ~ 1.5 orders of magnitude better than ELTL, meaning in particular that it can handle larger target queries.

Since Yago has only a very restricted ontology that essentially consists of inclusions $A \sqsubseteq B$ with A, B concept names, we complement the above experiment with a second one based on OWL2Bench. OWL2Bench is a benchmark for ontology-mediated querying that combines a database generator with a hand-crafted ontology which extends the University Ontology Benchmark [20, 21]. The ontology is formulated in OWL 2 EL and we extracted its \mathcal{ELH}^r fragment which uses all aspects of this DL and comprises 142 concept names, 83 role names, and 173 concept inclusions. We use datasets that contain 2500-2600 individuals and 100-200 examples, generated as in the Yago case. We designed 6 \mathcal{EL} -concepts with 3-5 occurrences of concept and role names and varying topology. The results are shown in Table 1. The difference in running time is even more pronounced in this experiment, with SPELL returning a fitting \mathcal{EL} -concept almost instantaneously in all cases.

We also performed initial experiments to evaluate how well the constructed fittings generalize to unseen data. We again use the Yago benchmark, but now split the examples into training data and testing data (assuming a uniform probability distribution). Table 2 lists the median accuracies of returned fittings (over 20 experiments) where the number of examples in the training data ranges from 5 to 65. As expected, fittings returned by SPELL generalize extremely well, even when the number of training examples is remarkably small. To our surprise, ELTL exhibits the same characteristics. This may be due to the fact that some heuristics of ELTL prefer fittings of smaller size, which might make ELTL an Occam algorithm. It would be interesting to carry out more extensive experiments on this aspect.

In [7], we carry out additional experiments in which we aim to highlight the respective strengths and weaknesses of SPELL and ELTL or, more generally, of bounded fitting versus refinement-operator based approaches. They show that the performance of bounded fitting is most affected by the number of existential restrictions in the target concept whereas the performance of refinement is most affected by the distance that the target concept has from \top in the subsumption lattice.

References

- [1] L. Bühmann, J. Lehmann, P. Westphal, DL-Learner - A framework for inductive learning on the semantic web, *J. Web Sem.* 39 (2016) 15–24.
- [2] N. Fanizzi, G. Rizzo, C. d’Amato, F. Esposito, DLFoil: Class expression learning revisited, in: *Proc. of EKAW*, 2018, pp. 98–113.
- [3] L. Iannone, I. Palmisano, N. Fanizzi, An algorithm based on counterfactuals for concept learning in the semantic web, *Appl. Intell.* 26 (2007) 139–159.
- [4] J. Lehmann, P. Hitzler, Concept learning in description logics using refinement operators, *Mach. Learn.* 78 (2010) 203–250.
- [5] M. Funk, J. C. Jung, C. Lutz, H. Pulcini, F. Wolter, Learning description logic concepts: When can positive and negative examples be separated?, in: *Proc. of IJCAI*, 2019, pp. 1682–1688.
- [6] J. C. Jung, C. Lutz, H. Pulcini, F. Wolter, Separating data examples by description logic concepts with restricted signatures, in: *Proc. of KR*, 2021, pp. 390–399.
- [7] B. ten Cate, M. Funk, J. C. Jung, C. Lutz, SAT-based PAC learning of description logic concepts, in: *Proc. of IJCAI*, 2023.
- [8] B. ten Cate, M. Funk, J. C. Jung, C. Lutz, SAT-based PAC learning of description logic concepts, *CoRR abs/2305.08511* (2023). URL: <https://arxiv.org/abs/2305.08511>.
- [9] A. Biere, A. Cimatti, E. M. Clarke, Y. Zhu, Symbolic model checking without BDDs, in: *Proc. of TACAS*, Springer, 1999, pp. 193–207.
- [10] A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *J. ACM* 36 (1989) 929–965.
- [11] L. G. Valiant, A theory of the learnable, *Commun. ACM* 27 (1984) 1134–1142.
- [12] R. A. Board, L. Pitt, On the necessity of Occam algorithms, *Theor. Comput. Sci.* 100 (1992) 157–184. doi:10.1016/0304-3975(92)90367-0.
- [13] M. Funk, Concept-by-Example in \mathcal{EL} Knowledge Bases, Master’s thesis, University of Bremen, 2019.
- [14] B. ten Cate, M. Funk, J. C. Jung, C. Lutz, On the non-efficient PAC learnability of acyclic conjunctive queries, *CoRR abs/2208.10255* (2022). doi:10.48550/arXiv.2208.10255. arXiv:2208.10255.
- [15] J. Kietz, Some lower bounds for the computational complexity of inductive logic programming, in: *Proc. of ECML*, 1993, pp. 115–123.
- [16] B. ten Cate, V. Dalmau, M. Funk, C. Lutz, Extremal fitting problems for conjunctive queries, in: *Proc. of PODS*, 2023.
- [17] B. Zarrieß, A. Turhan, Most specific generalizations w.r.t. general \mathcal{EL} -TBoxes, in: *Proc. of IJCAI*, 2013, pp. 1191–1197.
- [18] J. C. Jung, C. Lutz, F. Wolter, Least general generalizations in description logic: Verification and existence, in: *Proc. of AAAI*, AAAI Press, 2020, pp. 2854–2861.
- [19] T. P. Tanon, G. Weikum, F. M. Suchanek, YAGO 4: A reason-able knowledge base, in: *Proc. of ESWC*, Springer, 2020, pp. 583–596. doi:10.1007/978-3-030-49461-2_34.
- [20] G. Singh, S. Bhatia, R. Mutharaju, OWL2Bench: A benchmark for OWL 2 reasoners, in: *Proc. of ISWC*, Springer, 2020, pp. 81–96. doi:10.1007/978-3-030-62466-8_6.
- [21] Y. Zhou, B. C. Grau, I. Horrocks, Z. Wu, J. Banerjee, Making the most of your triple store:

query answering in OWL 2 using an RL reasoner, in: WWW, ACM, 2013, pp. 1569–1580.
doi:10.1145/2488388.2488525.