Concept Induction in Description Logics Using Information-Theoretic Heuristics

Nicola Fanizzi, University of Bari, Italy

ABSTRACT

This paper presents an approach to ontology construction pursued through the induction of concept descriptions expressed in Description Logics. The author surveys the theoretical foundations of the standard representations for formal ontologies in the Semantic Web. After stating the learning problem in this peculiar context, a FOIL-like algorithm is presented that can be applied to learn DL concept descriptions. The algorithm performs a search through a space of candidate concept definitions by means of refinement operators. This process is guided by heuristics that are based on the available examples. The author discusses related theoretical aspects of learning with the inherent incompleteness underlying the semantics of this representation. The experimental evaluation of the system DL-FOIL, which implements the learning algorithm, was carried out in two series of sessions on real ontologies from standard repositories for different domains expressed in diverse description logics.

Keywords: Concept Descriptions, Description Logics, FOIL Algorithms, Learning Algorithms, Ontology Construction, Semantic Web

1 INTRODUCTION

Formal ontologies are likely to play a key role in the next generation information systems moving from legacy to (linked) open data whose semantics is intended to be formalized and shared across the Web (Staab & Studer, 2009). One of the bottlenecks of this process is certainly represented by the construction (and evolution) of the ontologies since it involves different actors: domain experts contribute their knowledge but this is to be formalized by knowledge engineers so that it can be mechanized for the machines.

As the gap between these roles likely makes the process slow and burdensome, this problem may be tackled by resorting to machine learning techniques. Ontology learning (Cimiano, Mädche, Staab, & Völker, 2009) is intended to provide solutions to the problem of (semi-) automated ontology construction. Cast as an information extraction subtask, ontology learning has focused on learning from text corpora (Buitelaar & Cimiano, 2008). The main drawback of this approach is that the elicited concepts and relations are represented with languages of limited expressiveness. A different approach can be based on relational learning (see De Raedt, 2008, for a recent survey), which
requires a limited effort from domain experts (labeling individual resources as instances or non instances of the target concepts) and which leads to the construction of concepts even in very expressive languages (Lehmann, 2010).

If the concept learning problem is tackled as a search through a space of candidate descriptions in the reference representation guided by exemplars of the target concepts, then the same algorithms can be adapted to solve also ontology evolution problems. Indeed, while normally the semantics of change operations for this task has been considered from the logical and deductive point of view of automated reasoning, a relevant part of information lying in the data that populates ontological knowledge bases is generally overlooked or plays a secondary role.

Description Logics (DLs) is a family of languages supporting the standard ontology languages designed for knowledge bases in the context of the Semantic Web. These logics constitute specific fragments of First Order Logic (FOL) that differ from the standard clausal languages employed in relational learning, namely they have a different syntax and especially very different semantics (Borgida, 1996; Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2007). This motivates the growing interest in investigating inductive methods for such new formalisms.

1.1 Related Work

Early work on the application of machine learning to DLs essentially focused on demonstrating the PAC-learnability for various terminological languages derived from CLASSIC. In particular, Cohen and Hirsh investigate the CoreCLASSIC DL proving that it is not PAC-learnable (Cohen & Hirsh, 1992) as well as demonstrating the PAC-learnability of peculiar classes among its sub-languages, such as C-CLASSIC (Cohen & Hirsh, 1994), through the bottom-up LCSLEARN algorithm starting from lifted representatives of the instances obtained with another language-specific operator.

These approaches tend to cast supervised concept learning as performed through a structural generalizing operator working on equivalent graph representations of the concept descriptions. It is also worth mentioning unsupervised learning methodologies for DL concept descriptions, whose prototypical example is KLUSTER (Kietz & Morik, 1994), a polynomial-time algorithm for the induction of BACK terminologies, which exploits the tractability of the standard inferences in this DL language (Baader, Calvanese, et al., 2007).

More recently, approaches have been proposed that adopt the idea of generalization as search (Mitchell, 1982) performed through suitable operators that are specifically designed for DL languages (Badea & Nienhuys-Cheng, 2000; Fanizzi, Esposito, Ferilli, & Semeraro, 2003; Fanizzi, Ferilli, Iannone, Palmisano, & Semeraro, 2005; Esposito, Fanizzi, Iannone, Palmisano, & Semeraro, 2004; Iannone, Palmisano, & Fanizzi, 2007; Lehmann, 2010; Lehmann & Hitzler, 2010) on the grounds of the previous experience in the context of ILP which have been implemented in supervised (resp., unsupervised) learning systems, such as YIN-YANG (Iannone et al., 2007) and DL-LEARNER (Lehmann, 2009; Lehmann, Auer, Bühmann, & Tramp, 2011) (resp., CSKA) (Fanizzi, Iannone, Palmisano, & Semeraro, 2004). In particular, the OCEL algorithm (Lehmann & Hitzler, 2010) contained in the new releases of DL-LEARNER is able to learn class expressions in very expressive DL languages.

Learning alternative models such as logical decision trees offers another option for concept induction. The introduction of terminological decision trees (Fanizzi, d’Amato, & Esposito, 2010), i.e., logical decision trees with test-nodes represented by DL concept descriptions, and algorithms for their conversion into disjunctive descriptions, allows the combination of a divide-and-conquer learning strategy together with the standard separate-and-conquer strategy followed by most of the algorithms mentioned above (Boström & Asker, 1999).

In other cases, combined refinement operators have been proposed for learning in hybrid representations, integrating clausal and description logics (Grosof, Horrocks, Volz, &
Querying the Web Reconsidered: Design Principles for Versatile Web Query Languages
[www.igi-global.com/article/querying-web-reconsidered/2806?camid=4v1a](www.igi-global.com/article/querying-web-reconsidered/2806?camid=4v1a)

Managing Large Amounts of Data Generated by a Smart City Internet of Things Deployment