Ontology Learning from Text: Why the Ontology Learning Layer Cake is not Viable

Abel Browarnik, Tel Aviv University, Tel Aviv, Israel
Oded Maimon, Tel Aviv University, Tel Aviv, Israel

ABSTRACT

The goal of Ontology Learning from Text is to learn ontologies that represent domains or applications that change often. Manually learning and updating such ontologies is too expensive. This is the reason for the Ontology Learning discipline's emergence. The leading approach to Ontology Learning from Text is the Ontology Learning Layer Cake. This approach splits the task into four or five sequential tasks. Each of the tasks may use diverse methods, ranging from uses of Linguistic knowledge to Machine Learning. The authors review the shortcomings of the Ontology Learning Layer Cake approach and conclude that the approach is not viable for Ontology Learning from Text. They suggest alternative approaches that may help learning ontologies in an efficient, effective way.

Keywords: Deep Parsing, Entity-Relationship Diagrams, Linguistics based Ontology Learning, Ontology Learning from Text, Ontology Learning Layer Cake

INTRODUCTION

An ontology is a formal, explicit specification of a shared conceptualization. The ontology should be machine readable, with explicitly defined types of concepts and constraints on their use are. It should capture consensual knowledge, that is, it is not private to some individual, but accepted by a group.

Ontologies are often created and updated with human intervention. They represent reality, and therefore require frequent updates. Both the creation and the update of ontologies are costly activities. To overcome this problem the discipline of Ontology Learning has emerged. In this paper we refer specifically to Ontology Learning from text.

Surveys conducted since the early days of Ontology Learning show the different methods used to tackle the problem. The majority of methods follow an approach named the Ontology Learning Layer Cake. This approach splits the task into multiple steps towards learning an ontology, namely term extraction, concept formation, creation of a taxonomy of concepts, relation extraction and finally rules extraction.

This strategy does not take into account the conditions of the problem, nor the quality of the results obtained using the Ontology Learning Layer Cake methods. Working with well-formed
text (i.e., books, edited journal and other quality sources) is not the same as working with potentially lower quality sources, such as emails.

Moreover, splitting a task into smaller, sequential tasks, often help reduce complexity without undermining the results. As this paper shows, splitting the tasks, as in Ontology Learning Layer Cake methods, results in low quality results, making the methods unviable.

The paper is organized as follows:

• This introduction
• A review of the literature regarding Ontology Learning from text
• The target of Ontology Learning
• The input used for the Ontology Learning task
• The Ontology Learning Layer Cake approach
• Alternatives approaches
• Discussion and conclusion

ONTOLOGY AND ONTOLOGY LEARNING: A REVIEW

Ontologies

Gruber (1973) defines an ontology as an explicit specification of a conceptualization. The term is borrowed from philosophy, where ontology is a systematic account of what exists. For knowledge-based systems, what “exists” is exactly that which can be represented. Studer, Benjamins & Fensel (1998) define it as a formal, explicit specification of a shared conceptualization. A conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine readable, which excludes natural language. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.

Buitelaar, Cimiano, Grobelnik & Sintek (2005) classify ontologies following their use:

• Ontologies referring to high level concepts such as time, space, and event are classified as “top-level” ontologies and are generally domain independent. It makes sense to speak about unified, top-level ontologies.
• Ontologies describing the vocabulary related to a generic domain by specializing the concepts introduced in the top-level ontology are “domain ontologies”.
• A description of the vocabulary related to a generic task or activity by specializing the top-level ontologies is a “task ontology”.
• Finally, there are application ontologies. These are the most specific ontologies. Concepts in application ontologies often correspond to roles played by domain entities while performing a certain activity.

Top level ontologies, also called “upper” or “foundation” ontologies, were most likely constructed, at least initially, with human intervention. This is evidenced by Shah et al. (2006) for Cyc, Masolo, Borgo, Gangemi, Guarino & Oltramari (2003) for DOLCE, and Niles and Pease (2001) for SUMO. It is worth noting that there are only a few upper ontologies and their rate of change is allegedly low.