Towards an Error-Tolerant Construction of $\mathcal{EL}^\perp$-Ontologies from Data Using Formal Concept Analysis

Daniel Borchmann*

TU Dresden

Abstract. In the work of Baader and Distel, a method has been proposed to axiomatize all general concept inclusions (GCIs) expressible in the description logic $\mathcal{EL}^\perp$ and valid in a given interpretation $\mathcal{I}$. This provides us with an effective method to learn $\mathcal{EL}^\perp$-ontologies from interpretations. In this work, we want to extend this approach in the direction of handling errors, which might be present in the data-set. We shall do so by not only considering valid GCIs but also those whose confidence is above a given threshold $c$. We shall give the necessary definitions and show some first results on the axiomatization of all GCIs with confidence at least $c$. Finally, we shall provide some experimental evidence based on real-world data that supports our approach.

Keywords: Formal Concept Analysis, Description Logics, Ontology Learning.

1 Introduction

Description logic ontologies provide a practical yet formally well-defined way of representing large amounts of knowledge. They have been applied especially successfully in the area of medical and biological knowledge, one example being SNOMED CT [13], a medical ontology used to standardize medical nomenclature.

A part of description logic ontologies, the so called TBox, contains the terminological knowledge of the ontology. Terminological knowledge constitutes connections between concept descriptions and is represented by general concept inclusions (GCIs). For example, we could fix in an ontology the fact that everything that has a child is actually a person. Using the description logic $\mathcal{EL}^\perp$, this could be written as

$$\exists\text{child.} \top \sqsubseteq \text{Person}.$$  

Here, $\exists\text{child.}$ and Person are examples of concept descriptions, and the $\subseteq$ sign can be read as “implies.” General concept inclusions are, on this intuitive level, therefore quite similar to implications.

The construction of TBoxes of ontologies, which are supposed to represent the knowledge of a certain domain of interest, is normally conducted by human experts. Although this guarantees a high level of quality of the resulting ontology, the...
process itself is long and expensive. Automating this process would both decrease the time and cost for creating ontologies and would therefore foster the use of formal ontologies in other applications. However, one cannot expect to entirely replace human experts in the process of creating domain-specific ontologies, as these experts are the original source of this knowledge. Hence constructing ontologies completely automatically does not seem reasonable.

A compromise for this would be to devise a semi-automatic way of constructing ontologies, for example by learning relevant parts of the ontology from a set of typical examples of the domain of interest. The resulting ontologies could be used by ontology engineers as a starting point for further development.

This approach has been taken by Baader and Distel [8,2] for constructing $\mathcal{EL}/C^3$-ontologies from finite interpretations. The reason why this approach is restricted to $\mathcal{EL}/C^3$ is manifold. Foremost, this approach exploits a tight connection between the description logic $\mathcal{EL}/C^3$ and formal concept analysis [9], and such a connection has not been worked out for other description logics. Moreover, the description logic $\mathcal{EL}/C^3$ can be sufficient for practical applications, as, for example, SNOMED CT is formulated in a variant of $\mathcal{EL}/C^3$. Lastly, $\mathcal{EL}/C^3$ is computationally much less complex than other description logics, say $\mathcal{ALC}$ or even $\mathcal{FL}_0$.

In their approach, Baader and Distel are able to effectively construct a base of all valid GCIs of a given interpretation, where this interpretation can be understood as the collection of typical examples of our domain of interest. This base therefore constitutes the complete terminological knowledge that is valid in this interpretation. Moreover, these interpretations can be seen as a different way to represent linked data [3], the data format used by the semantic web community to store its data. Hence, this approach allows us to construct ontologies from parts of the linked data cloud, providing us with a vast amount of real-world data for experiments and practical applications.

In [7], a sample construction has been conducted on a small part of the DBpedia data set [4], which is part of the linked open data cloud. As it turned out, the approach is effective. However, another result of these experiments was the following observation: in the data set extracted from DBpedia, a small set of errors were present. These errors, although very few, greatly influenced the result of the construction, in the way these errors invalidated certain GCIs, and hence these GCIs were not extracted by the algorithm anymore. Then, instead of these general GCIs, more special GCIs were extracted that “circumvent” these errors by being more specific. This not only lead to more extracted GCIs, but also to GCIs which may be hard to comprehend.

As the original approach by Baader and Distel considers only valid GCIs, even a single error may invalidate an otherwise valid GCI. Since we cannot assume from real-world data that it does not contain any errors, this approach is quite limited for practical applications. Therefore, we want to present in this work a generalization to the approach of Baader and Distel which does not only consider valid GCIs but also those which are “almost valid.” The rationale behind this is that these GCIs should be much less sensitive to a small amount of errors than valid GCIs. To decide whether a GCI is “almost valid,” we shall use its confidence in the given