Master Thesis Topic: Extracting Ontologies from Large Language Models with Queries and Counterexamples

Supervisor: Ana Ozaki
University of Oslo, Norway

1 Introduction and Motivation

Language models such as GPT have been trained on large amounts of data and can generate text in various domains of knowledge. However, these models lack the ability of performing logical reasoning [6]. Ontologies, on the other hand, are useful for representing the relevant knowledge of a domain of interest in a format that allows for logical reasoning in an automated way [2]. Though, building ontologies is known to be an expensive and time-consuming task.

Example 1 *To illustrate how an ontology can look like, consider the following ontology, written in the classical $\mathcal{EL}$ description logic language, about family relations:*

$$\{\exists\mathsf{hasChild}.\top \sqsubseteq \mathsf{Parent}, \mathsf{Woman} \sqcap \exists\mathsf{sibling}.\mathsf{Parent} \sqsubseteq \mathsf{Aunt}\}$$

which expresses that (1) if someone has a child then it is a parent and (2) if a woman has a sibling who is a parent then she is an aunt. This means that if we have the data

$$\{\mathsf{Woman}(\mathsf{Mary}), \mathsf{sibling}(\mathsf{Mary}, \mathsf{John}), \mathsf{hasChild}(\mathsf{John}, \mathsf{Alice})\}$$

which expresses that (1) Mary is a woman, (2) Mary and John are siblings and (3) John has child who is Alice, then, by logical reasoning, one can deduce from the ontology and the data that John *is a parent and* Mary *is an aunt.*

In this project, the candidate will explore the possibility of extracting an ontology from a language model by posing queries in Angluin's exact learning framework [1, 5]. In the exact learning framework, a learner attempts to extract information from a teacher by posing queries (which represent questions). The idea in this project is to see the language model as the teacher and apply an algorithm designed for learning $\mathcal{EL}$ ontologies via queries and counterexamples [4] to this setting.

2 Research questions

The main research question is whether and how one can extract an ontology from a language model by posing queries. This could help ontology engineers by providing them a draft of an ontology created in this automated way, as well as possibly detecting harmful biases and/or false information in language models. We propose to consider the $\mathcal{EL}$ description logic ontology language, which has been extensively studied in the field of knowledge representation and allows for good expressivity power while retaining polynomial time complexity for many logical reasoning tasks [2]. Even though the algorithm itself is independent of the domain, it is important to limit the scope. For simplicity, we suggest as a first domain option building an ontology about family relations and then performing additional experiments in some other domains.

3 Thesis Work

In this master thesis topic, the candidate of a master degree will study an algorithm from the literature and apply it for systematically probing language models. This is the algorithm for extracting $\mathcal{EL}$ ontologies with queries and counterexamples by Duarte et al. [4]. The main tasks of this project are

- to study reference papers from the literature on the topic of the project (suggested by the supervisor and found by the candidate),
- to study how to apply the algorithm for exact learning $\mathcal{EL}$ ontologies to extract knowledge in a domain of interest from a language model, following the initial steps by Duarte et al. and Blum et al. [4, 3], and
- to implement the mentioned algorithm (suggested programming language: python or Java) and to apply it for probing language models in a systematic way.

It is also expected that the candidate performs experiments showcasing the applicability of the implementation. As already mentioned, it is suggested that, in the experiments, the candidate attempts to extract an ontology about family relations using language models such as GPT. It would be interesting to perform additional experiments using other domains (to be decided together with the candidate) and check whether and how the ontologies extracted resemble already existing ontologies built manually by ontology engineers in the conventional way.

References

[1] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.

[2] Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, second edition, 2007.

[3] Sophie Blum, Raoul Koudijs, Ana Ozaki, and Samia Touileb. Learning horn envelopes via queries from large language models. *CoRR*, abs/2305.12143, 2023.

[4] Mario Ricardo Cruz Duarte, Boris Konev, and Ana Ozaki. Exactlearner: A tool for exact learning of EL ontologies. In Michael Thielscher, Francesca Toni, and Frank Wolter, editors, *KR*, pages 409–414. AAAI Press, 2018.

[5] Boris Konev, Carsten Lutz, Ana Ozaki, and Frank Wolter. Exact learning of lightweight description logic ontologies. *J. Mach. Learn. Res.*, 18:201:1–201:63, 2017.

[6] Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. *CoRR*, abs/2205.11502, 2022.