Actively Learning Ontologies from LLMs: First Results

Matteo Magnini, Ana Ozaki, Riccardo Squarcialupi

matteo.magnini@unibo.it, anaoz@uio.no, riccard.squarcialupi@studio.unibo.it

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA · UNIVERSITY OF BERGEN · UNIVERSITY OF OSLO

Actively Learning & Vision

In active learning a **learner** attempts to learn some kind of knowledge by posing questions to a **teacher**.

Questions made by the learner are called **membership queries** and are answered with **yes/true** or **no/false**.

Vertebrate SubClass of Animal?
True
Mammal SubClass of Vertebrate?
True
Carnivore SubClass of Vertebrate?
False

We consider the case in which the knowledge is expressed as an $\mathcal{EL}$ **ontology**. Membership queries consist in asking if an axiom belongs to the ontology.

Our intention is to first use a **large language model** (LLM) as a teacher for actively learning ontologies and evaluate the results.

The Angluin's **Exact Learning** framework makes use of active learning when membership queries are allowed.

Currently, the only implementation for learning $\mathcal{EL}$ ontologies in the exact learning framework is with a **synthetic teacher**, created by the authors for testing the implementation.

Right now, we are working on an extension of the **ExactLearner** [1] to use LLMs as teachers.

Probing Language Models

Challenges

- **Input format:** questions standardisation to systematically query an LLM. We investigate the use of the *Manchester OWL syntax* (rigorous formalism and close to natural language).

- **Unexpected responses:** LLMs may answer with an arbitrary response. We use a custom *system prompt* and we set a fixed maximum number of *tokens* to mitigate this issue. A post processing phase to handle the response is still needed.

- **Correctness & logical consistency:** there is no guarantee that the responses are correct (i.e., true in the real world). Moreover, they may not be logically consistent. For example, all concept inclusions in $\mathcal{T} = \{C1 \sqsubseteq D1, \ldots, Cn \sqsubseteq Dn\}$ are answered with true, but $\mathcal{T} \models C \sqsubseteq D$ but $C \sqsubseteq D$ is classified as false.

↓

We search for logical inconsistency by creating the closure under logical consequence and testing whether something in the closure received **false** as answer. Therefore, in the previous example we consider $C \sqsubseteq D$ as true.

Findings

- **Some inconsistencies:** we observed and measured logical inconsistencies in the responses of the LLMs;

- **Good performace:** overall, there is statistical evidence that the answers of the LLMs (in particular *GPT 3.5 Turbo*, *Mistral* and *Mixtral*) correlate with the knowledge in the ontologies.

Experimental Evaluation

Case Study

Perform a number of membership queries with multiple LLMs, without any fine-tuning, on $\mathcal{EL}$ ontologies. Experiments:

1. check how well LLMs answer to membership queries using the logical axioms in the ontologies;

2. we repeat the experiments in 1, but using the inferred axioms (the logical closure of the ontologies we use is finite!);

3. we actively learn ontologies by means of a naive algorithm where all concept inclusions of the form $A \sqsubseteq B$ with A, B concept names in a given signature are asked.

Ontologies

We use five ontologies used for experiments in the ExactLearner project [1]: *Animals*, *Cell*, *Football*, *Generation* and *University*.

LLMs

We choose five LLMs: *GPT 3.5 Turbo* (?b), *Mistral* (7b), *Mixtral* (47b), *Llama 2* (7b) and *LLama 2* (13b).

Metrics

For experiments 1 and 2 we compute the number of true, false and unknown (i.e., neither true nor false) answers. In experiment 2 we also report the logical inconsistencies found. Note that because these axioms are present in the ontologies an LLM that does not make mistakes must reply with true.

For experiment 3 we report accuracy, precision and recall. The axioms used for membership queries are both present and not present in the ontologies.

[1] M. R. C. Duarte, B. Konev, A. Ozaki, Exactlearner: A tool for exact learning of $\mathcal{EL}$ ontologies, in KR 2018.

Results

Models	Animals			University			Generations			Football			Cell		
	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U
Mistral (7b)	9	1	2	2	0	2	5	10	3	7	2	0	17	1	6
Mixtral (47b)	11	1	0	4	0	0	3	6	9	9	0	0	15	9	0
Llama2 (7b)	11	1	0	4	0	0	16	1	1	9	0	0	24	0	0
Llama2 (13b)	11	1	0	4	0	0	16	1	1	9	0	0	23	1	0
Gpt3.5	10	2	0	4	0	0	13	4	1	9	0	0	21	3	0

Table 1
Results for the experiments testing correctness w.r.t. axioms in the ontologies. Labels T, F and U mean true, false and unknown responses count.

Animals				University				Generations				Football				Cell			
T	F	U	L	T	F	U	L	T	F	U	L	T	F	U	L	T	F	U	L
14	2	4	2	5	1	2	0	10	27	5	2	9	3	0	0	18	1	5	0
18	2	0	0	8	0	0	0	19	13	10	0	12	0	0	0	17	7	0	0
20	0	0	0	8	0	0	0	40	1	1	1	12	0	0	0	24	0	0	0
18	2	0	1	7	1	0	0	35	6	1	4	11	1	0	1	21	3	0	0
20	0	0	0	7	1	0	0	36	5	1	0	12	0	0	0	18	6	0	0

Table 2
Results for the experiments testing logical consistency. The meaning of T, F and U is the same as in Table 1. L stands for logical inconsistencies.

Animals			University			Generations			Football			Cell		
A	P	R	A	P	R	A	P	R	A	P	R	A	P	R
0.87	0.52	0.72	0.57	0.67	0.5	0.84	0.71	0.23	0.74	0.44	0.65	0.65	0.48	0.81
0.89	0.57	0.69	0.57	0.48	0.92	0.82	0.64	0.66	0.72	0.43	0.76	0.7	0.32	0.64
0.51	0.2	1	0.24	0.24	1	0.4	0.22	0.88	0.21	0.21	1	0.27	0.18	1
0.73	0.31	0.94	0.45	0.3	0.92	0.63	0.32	0.74	0.44	0.26	0.88	0.44	0.21	0.91
0.71	0.3	1	0.69	0.44	1	0.74	0.41	1	0.68	0.4	1	0.61	0.28	0.91

Table 3
Results for the experiments testing negative examples. Labels A, P and R mean Accuracy, Precision and Recall. We applied the Chi-squared test to check the relationship between the answers of the LLMs and the ontologies, with the null hypothesis being that there is no correlation (yellow cells).

Link to the github repository here!!!