Project SHADOW: Symbolic Higher-order Associative Deductive reasoning On Wikidata using LM probing

Hanna Abi Akl

Data ScienceTech Institute, Université Côte d'Azur, Inria, CNRS, I3S, Paris, France

Abstract. We introduce SHADOW, a fine-tuned language model trained on an intermediate task using associative deductive reasoning, and measure its performance on a knowledge base construction task using Wikidata triple completion. We evaluate SHADOW on the LM-KBC 2024 challenge and show that it outperforms the baseline solution by 20% with a F1 score of 68.72%.

Keywords: Knowledge graphs, ontologies, natural language processing, large language models.

1 Introduction

Large language models (LLMs) have performed increasingly well in a wide range of semantic tasks including those involving leveraging knowledge from the models themselves [11]. This lead to research avenues investigating the capabilities of these models in knowledge-related tasks involving knowledge graphs and ontologies on the one hand, and measuring the intrinsic knowledge contained in LLMs on the other hand [11]. The Language Model Knowledge Base Construction (LM-KBC[1]) challenge proposes to evaluate intrinsic language model (LM) knowledge using techniques like LM probing and prompting [8] to construct knowledge bases by completing triples of subject entities and relations with the relevant object entities. In this work, we present SHADOW, or **S**ymbolic **H**igher-order **A**ssociative **D**eductive reasoning **O**n **W**ikidata, a fine-tuned model on knowledge base triples, and evaluate it on the LM-KBC task. We follow a methodology inspired from associative deductive reasoning [9] and leverage that technique to incorporate it in re-defining the probing problem to train the model more effectively. Specifically, we aim to address the following research questions: Can LLMs use transfer learning as a means to leverage their generative abilities to understand an association task? Can they use deductive reasoning to solve it based on a neighboring task they have been previously trained on? The rest of the work is organized as follows. In section 2, we discuss some of the related work. Section 3 describes our experimental framework. In section 4, we report our results and discuss our findings. Finally, we conclude in section 5.

2 Related work

LM probing has been studied and evaluated in different research avenues. In their work, [10] study the information stored in LLMs with respect to their architecture, focusing on the factors behind their understanding of lexical semantics. Other techniques leverage prompting to encourage LLMs to use their knowledge more effectively to find better answers. [1] show that curating prompts manually and combining them with sets of entities

[1] https://lm-kbc.github.io/challenge2024/

David C. Wyld et al. (Eds): NATL, SEAPP, SIGEM, MLTEC, CSEA – 2024
pp. 01-07, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142101

make relatively small LLMs perform well on knowledge base construction. Similar work shows that LLMs can be prompted to generate seemingly coherent responses to incoherent inputs[3].

Other lines of work treat LLMs as knowledge bases and employ query-based techniques to probe them for specific knowledge. [8] compare different transformer-based models by querying them with specific input knowledge and tracing whether this knowledge is retained in the LLMs. [2] evaluate LLMs as knowledge bases using different techniques ranging from model editing to discrete prompting on defined metrics like interpretability and causal tracing.

Research has also shown work on LLMs as knowledge bases using external knowledge in the form of extended vocabulary [6], knowledge sources like knowledge graph information [7] or by designing architectures that support external vectorized knowledge sources like Retrieval-Augmented-Generation (RAG) systems [5]. Finally, while deductive abilities for LLMs have been heavily investigated, [4] perform experiments to evaluate the inductive reasoning capabilities of these models and find that their deductive reasoning are much poorer than their inductive reasoning skills.

3 Experiments

This section describes our experiments in terms of data, model and training process.

3.1 Dataset

The data provided by the organizers are triples of the form (subject, relation, object). The following relations are considered:

- **countryLandBordersCountry:** Null values possible (e.g., Iceland)
- **personHasCityOfDeath:** Null values possible
- **seriesHasNumberOfEpisodes:** Object is numeric
- **awardWonBy:** Many objects per subject (e.g., 224 Physics Nobel prize winners)
- **companyTradesAtStockExchange:** Null values possible

The data is provided in 3 sets: train, validation and test. The test set is used as the official submission evaluation set. The number of triples in each set is:

- 377 in the train set
- 378 in the validation set
- 378 in the test set

For the subject and object in every triple, both the ID and the label are provided. A sample triple is thus represented as such: {*"SubjectEntity": "Belize", "SubjectEntityID": "Q242", "ObjectEntities": ["Guatemala", "Mexico"], "ObjectEntitiesID": ["Q774", "Q96"], "Relation": "countryLandBordersCountry"*}.

3.2 Model

Formally, SHADOW is a generative model designed to solve the following function: $y = f(x)$, where x is the input triple $(s, p, \cdot)$ missing object entities, f is the function that generates a number in the set $\{1, 2, 3, 4, 5\}$ and y is the corresponding template from the set of templates $T = \{t_1, t_2, t_3, t_4, t_5\}$, where t is a SPARQL query responsible for completing the knowledge graph triple with the missing objects. We train SHADOW as a conditional generation model from a base flan-t5-small[2] model and fine-tune it on the provided data. The training hyperparameters are configured as such:

– **learning_rate:** 1e-04
– **train_batch_size:** 4
– **eval_batch_size:** 4
– **num_epochs:** 20
– **question_length:** 512
– **target_length:** 512
– **lr_scheduler_type:** linear
– **optimizer:** Adam
– **betas:** 0.9, 0.999
– **epsilon:** 1e-08

3.3 Setup

We design our experiment to combine LLM probing with a symbolic component and indirectly evaluate the intrinsic knowledge found in LLMs on Wikidata knowledge graphs. We shift the focus away from generating correct SPARQL queries to retrieving the relevant objects for each subject and relation pair by designing templates containing the dynamic queries needed to answer the generic question:

What Z completes the relationship Y for X?,

where X, Y and Z refer respectively to the subject, relation and object(s) in a triple. Since the challenge deals with 5 types of relations, we design a total of 5 templates and assign a numerical template ID to each one of them. We then pair each subject and relation from a triple with the corresponding template ID which points to the correct SPARQL query that retrieves the corresponding object(s) to complete the triple based on the relation type. The SPARQL queries are themselves designed based on the Wikidata properties that best represent the targeted relation in the given triples.

SHADOW is then trained to generate the correct template ID depending on the given subject and relation without seeing the SPARQL queries. This requires the model learning on two fronts: first is by implicitly learning to associate the correct template with the relation type, and second by learning to generate a numerical value which corresponds to one of the acceptable template IDs defined in the frame of the experiment. Figure 1 shows our experimental design process.

The training is done by splitting the train set into an 80-20 split randomly and training on the 80%. The remaining 20% are incorporated into the validation set. We use the scikit-learn[3] library to perform the data splits. Table 1 shows the training results.

The experiment is conducted on a Google Colab instance using a L4 High-RAM GPU. The code for our experimental setup is publicly available on GitHub[4]. The SHADOW model has been publicly released on Hugging Face[5].

4 Results

Tables 2 and 3 capture the results on the official challenge test set. Overall, SHADOW performs well on the template identification task for the different relations. A closer inspection

[2] https://huggingface.co/google/flan-t5-small
[3] https://scikit-learn.org/stable/
[4] https://github.com/HannaAbiAkl/SHADOW
[5] https://huggingface.co/HannaAbiAkl/shadow

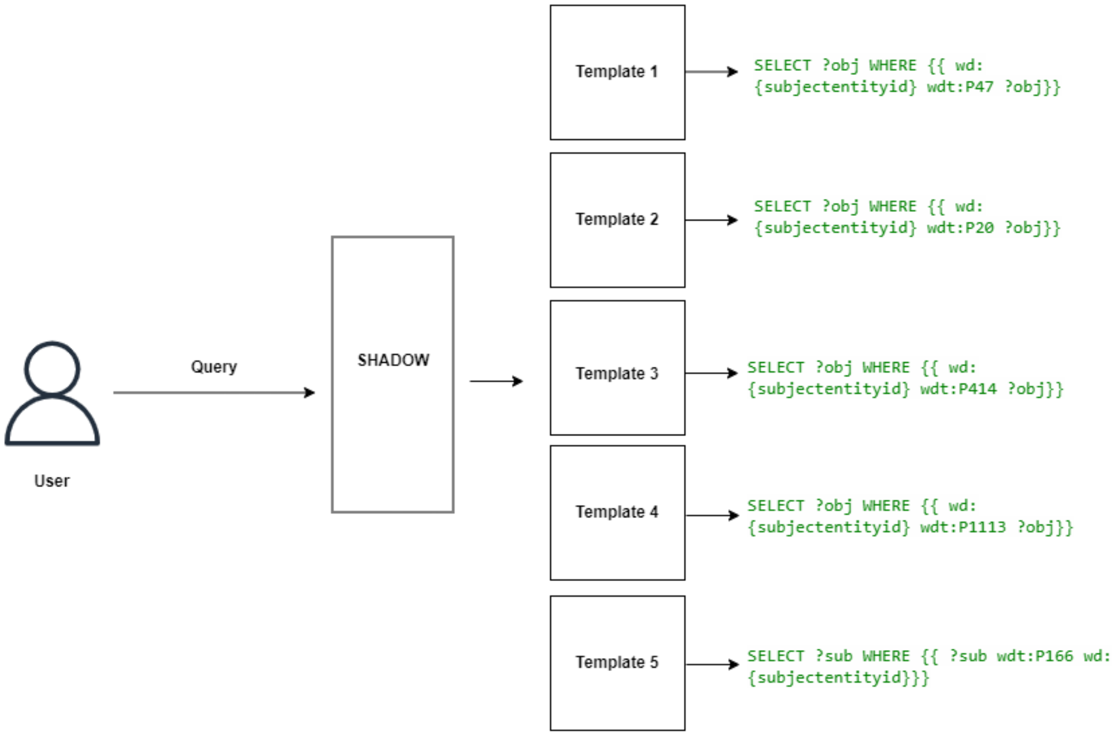

Fig. 1. Experimental setup.

Table 1. SHADOW model training results

Training Loss	Epoch	Step	Validation Loss
0.6679	1.0	1000	0.0005
0.3425	2.0	2000	0.0002
0.2303	3.0	3000	0.0001
0.1735	4.0	4000	0.000
0.1394	5.0	5000	0.0001
0.1167	6.0	6000	0.00009
0.1006	7.0	7000	0.00008
0.0882	8.0	8000	0.00007
0.0785	9.0	9000	0.00006
0.0707	10.0	10000	0.00006
0.0643	11.0	11000	0.00005
0.0590	12.0	12000	0.00005
0.0545	13.0	13000	0.00004
0.0506	14.0	14000	0.00004
0.0473	15.0	15000	0.00004
0.0443	16.0	16000	0.00003
0.0417	17.0	17000	0.00003
0.0394	18.0	18000	0.00003
0.0374	19.0	19000	0.00003
0.0355	20.0	20000	0.00003

of the results shows that the model performs worse on the *countryLandBordersCountry* relation which can be interpreted by the fact that the corresponding query targets property *P47*, or *shares border with*, which encompasses but is not limited to results sharing land borders (i.e. it also targets objects sharing sea borders). The nature of the property explains the high recall score, which retains a larger set of objects for that relation, and the low precision which reflects the few actual correct objects expected. The reason for choosing P47 is that it was the closest property that meets the required relation.

The result is validated in the zero-object cases, wherein the recall score is fairly high compared to a low precision score. This results is directly impacted by the choice of coding the queries as templates, sacrificing flexibility in results in favor of correct syntax instead of leaving the query generation in the hands of the model and risking some volatility in the results.

The relation *seriesHasNumberOfEpisodes* shows a contrasting result. The tradeoff between perfect precision and zero recall suggests a cautious classification, whereby the net of positive results, i.e. relations correctly associated with the proper template ID, is small but accurate. The precision score is explained by the fact that the query behind the template targets the correct property and returns the correct object (which is either a number or null if there are no series episodes). The low recall means SHADOW did not learn to associate this particular relation with its correct template, drawing questions over the intrinsic knowledge for that type of relation in the model. It is also possible that the stark difference between the nature of this relation, which is the only one among the 5 to expect a purely numerical answer as opposed to Wikidata entities, has proven more challenging for SHADOW to learn despite the fine-tuning process it has undergone.

Finally, it is worth highlighting the model's great performance on the *awardWonBy* relation, considering it is underrepresented at one-tenth of the other relations, i.e. 10 relations in the train, validation and test sets compared to approximately 100 for each of the four other relations. Since the model received less data for this relation compared to the others and still performed well, the performance could be explained by the nature of the internal knowledge the model has amassed for this relation, which in turn enables us to hypthesize on the good quality of the data given to this model at pre-training.

Table 2. Per-relation scores

Relation	Precision	Recall	F1-score
awardWonBy	0.9816	1.0000	0.9900
companyTradesAtStockExchange	0.9950	1.0000	0.9971
countryLandBordersCountry	0.7470	0.9717	0.7829
personHasCityOfDeath	0.9700	1.0000	0.9700
seriesHasNumberOfEpisodes	1.0000	0.0000	0.0000
Average	0.9453	0.7297	0.6872

Table 3. Zero-object cases

Precision	Recall	F1-score
0.4975	0.90006	0.6408

Table 4 shows the official performance of SHADOW compared to other solutions. Despite its difficulties with some relations, SHADOW performs very well with respect to the baseline and outperforms it by almost 20%. It falls however short of the best scores and is outclassed by some margin.

Table 4. Official submission leaderboard

Team Name	Average F1-score
davidebara	0.9224
KB	0.9131
RAGN4ROKS	0.9083
WWWD	0.6977
DSTI	**0.6872**
NadeenFathallah	0.6529
Rajaa	0.5662
aunsiels	0.5076
lm-kbc-organizer	0.4865

5 Conclusion

In this work, we show how a fine-tuned LLM model can leverage intrinsic knowledge through LM probing and combine it with associative deductive reasoning to build disambiguated knowledge bases. The performance of SHADOW, our model, outperforms the baseline by disambiguating relation types and indirectly associating them with relevant knowledge graph completion queries. Our experiments show however that LLMs are highly influenced by the type of data they have already been trained on and possess uneven knowledge with respect to Wikidata relations which leaves much room for improvement in that area. Future work will focus on studying relation types in depth to improve LLM knowledge and using that knowledge to further evaluate LLM reasoning capabilities.

References

1. Alivanistos, Dimitrios and Santamaría, Selene Báez and Cochez, Michael and Kalo, Jan-Christoph and van Krieken, Emile and Thanapalasingam, Thiviyan: Prompting as probing: Using language models for knowledge base construction. **arXiv preprint arXiv:2208.11057** (2022)
2. AlKhamissi, Badr and Li, Millicent and Celikyilmaz, Asli and Diab, Mona and Ghazvininejad, Marjan: A review on language models as knowledge bases. **arXiv preprint arXiv:2204.06031** (2022)
3. Cherepanova, Valeriia and Zou, James: Talking Nonsense: Probing Large Language Models' Understanding of Adversarial Gibberish Inputs. **arXiv preprint arXiv:2404.17120** (2024)
4. Cheng, Kewei and Yang, Jingfeng and Jiang, Haoming and Wang, Zhengyang and Huang, Binxuan and Li, Ruirui and Li, Shiyang and Li, Zheng and Gao, Yifan and Li, Xian and others: Inductive or deductive? Rethinking the fundamental reasoning abilities of LLMs. **arXiv preprint arXiv:2408.00114** (2024)
5. Li, Belinda Z and Liu, Emmy and Ross, Alexis and Zeitoun, Abbas and Neubig, Graham and Andreas, Jacob: Language Modeling with Editable External Knowledge. **arXiv preprint arXiv:2406.11830** (2024)
6. Loureiro, Daniel and Jorge, Alípio Mário: Probing Commonsense Knowledge in Pre-trained Language Models with Sense-level Precision and Expanded Vocabulary. **arXiv preprint arXiv:2210.06376** (2022)
7. He, Qiyuan and Wang, Yizhong and Wang, Wenya: Can Language Models Act as Knowledge Bases at Scale?. **arXiv preprint arXiv:2402.14273** (2024)

8. Petroni, Fabio and Rocktäschel, Tim and Riedel, Sebastian and Lewis, Patrick and Bakhtin, Anton and Wu, Yuxiang and Miller, Alexander: Language Models as Knowledge Bases?. **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp 2463–2473 (2019)
9. Takano, Wataru and Nakamura, Yoshihiko: Associative processes between behavioral symbols and a large scale language model. **2010 IEEE International Conference on Robotics and Automation**, pp 2404–2409 (2010)
10. Vulić, Ivan and Ponti, Edoardo Maria and Litschko, Robert and Glavaš, Goran and Korhonen, Anna: Probing pretrained language models for lexical semantics. **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp 7222–7240 (2020)
11. Wang, Xin and Chen, Zirui and Wang, Haofen and Li, Zhao and Guo, Wenbin and others: Large Language Model Enhanced Knowledge Representation Learning: A Survey. **arXiv preprint arXiv:2407.00936** (2024)

Authors

Hanna Abi Akl is a professor and researcher in logic, language, computation and neuro-symbolic artificial intelligence. He holds a Masters degree in Computer Science from the Université Côte d'Azur and an MSc in Data Science and Artificial Intelligence from Data ScienceTech Institute. His main areas of research are mapping language to symbolic structures, logic-based reasoning and graph-based knowledge representation methods. Hanna works as a researcher at INRIA and holds the Vice-Dean position at Data ScienceTech Institute.