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

Hanna Abi Akl

NATL 2024

Content

- Task Introduction
- Dataset
- Definition – Associative Learning
- Definition – Transfer Learning
- Definition – Associative Deductive Task Learning
- Model
- Experimental Setup
- Results
- Conclusion

Task Introduction

- Language Model Knowledge Base Construction challenge (LM-KBC)
- Proposed at ISWC 2024
- Formally: Given the input subject-entity (s) and relation (r), the task is to predict all the correct object-entities ({o1, o2, ..., ok}) using LM probing
- *RQ1: Can LLMs use deductive reasoning capabilities to understand a new task that shares the same dataset they have been trained on to solve another task?*
- *RQ2: How effectively can LLMs use intrinsic knowledge to solve a new task?*

Dataset

- Wikidata triples of the form (subject, relation, object)
- Limited to the following relations: *countryLandBordersCountry, personHasCityOfDeath, seriesHasNumberOfEpisodes, awardWonBy, companyTradesAtStockExchange*
- 377 triples in train set
- 378 triples in validation set
- 378 triples in test set

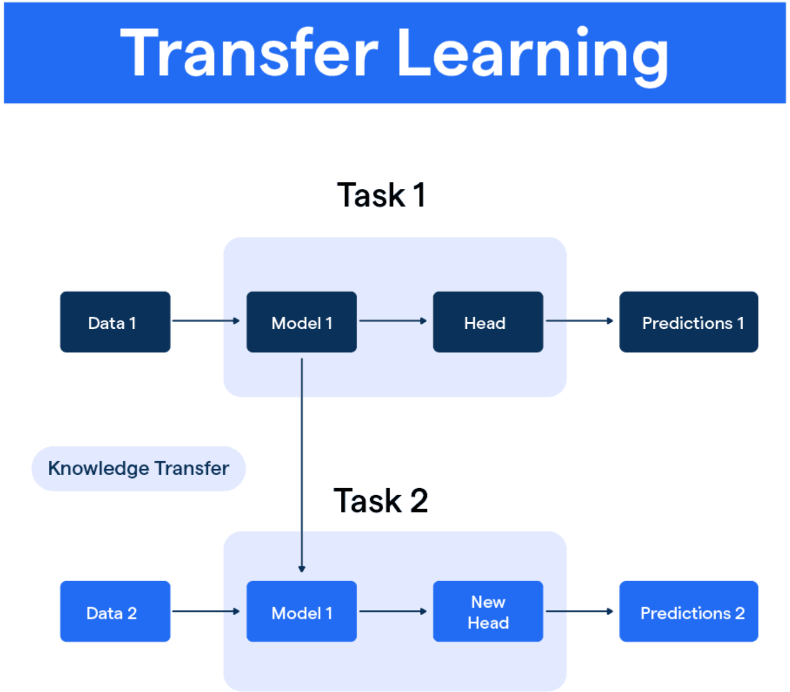
For the subject and object in every triple, both the ID and the label are provided. A sample triple is thus represented as such: {*”SubjectEntity”: ”Belize”, ”SubjectEntityID”: ”Q242”, ”ObjectEntities”: [”Guatemala”, ”Mexico”], ”ObjectEntitiesID”: [”Q774”, ”Q96”], ”Relation”: ”countryLandBordersCountry”*}.

Definition – Associative Learning

What is Associative Learning?

Associative learning is when two stimuli become linked or learned in tandem. The elements of one stimulus then become associated with the second stimulus.

Definition – Transfer Learning



Definition – Associative Deductive Task Learning (1)

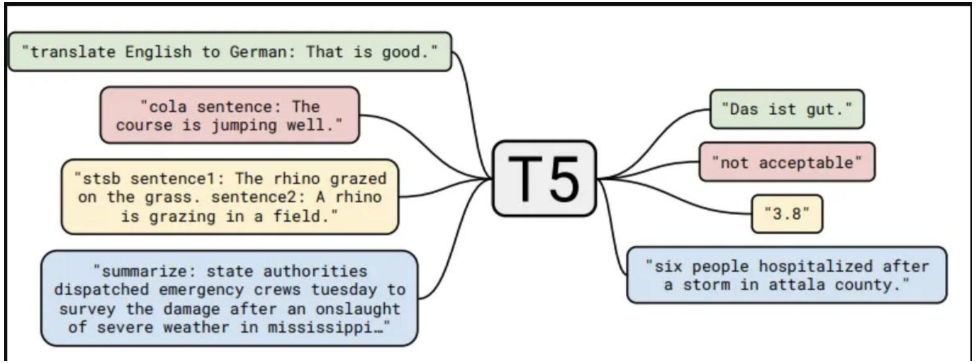
- Combines the best of both Associative Learning and Transfer Learning
- Designs a new (unseen) task for the LLM on a seen dataset
- Requires associative and deductive reasoning to solve task
- Requires use of key intrinsic knowledge in LLM
- Formally: Generate number in *set {1,2,3,4,5}* corresponding to *t* in set of *templates T = {t1,t2,t3,t4,t5}* where *t* is a *SPARQL query* for knowledge graph completion for each relation r in a Wikidata triple

Definition – Associative Deductive Task Learning (2)

- Why redefine the task?
  - Problem simplification: From graph completion to classification
  - No additional input (dataset) needed: relies on LLM intrinsic knowledge on seen data
  - Grounded problem: Templates are logical (limits hallucinations), human-readable (human-in-the-loop), transparent (explainable)

Model

- **S**ymbolic **H**igher-order **A**ssociative **D**eductive reasoning **O**n **W**ikidata
- Small Language Model (SLM): Fine-tuned on flan-t5-small
- Trained in a classification setting to generate (predict) best number of template
- Oblivious to the SPARQL query (answer) behind each template



NATL 2024

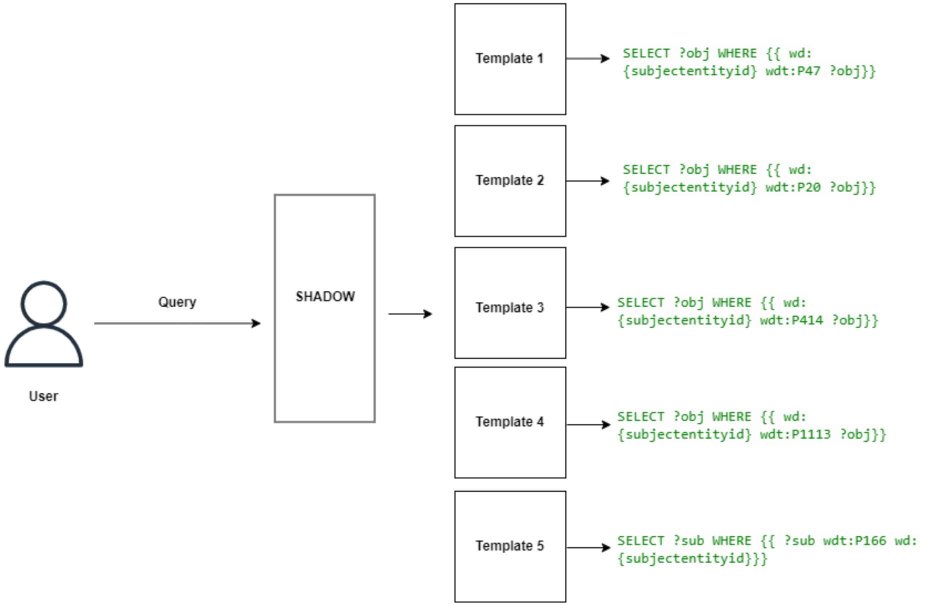
wimmics Inria INVENTEURS DU MONDE NUMÉRIQUE

UNIVERSITÉ CÔTE D'AZUR

DSTI

Experimental Setup

- Experiment designed in a question-answering (QA) setting
- Question prompt:
  - *What Z completes the relationship Y for X?*
  - *X = subject; Y = relation; Z = object(s)*
- SHADOW trained to generate correct template ID for the correct query given the subject and relation of a triple
- Any other output generated is considered faulty and incorrect
- 80% of train data used for training
- Remaining 20% added to validation set



NATL 2024

Results (1)

- *countryLandBordersCountry* bad performance due to SPARQL query which targets *P47* Wikidata property
- *seriesHasNumberOfEpisodes* scores suggest cautious classification
- Also only relation to expect numerical objects
- *awardWonBy* score shows valuable use of intrinsic knowledge since data samples represents 1/10 compared to other relations

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

Results (2)

- SHADOW outperforms baseline in LM-KBC task by 20%
- Falls a long way behind other systems
- Limitations suggest possible revision of intrinsic knowledge amassed by the base model

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

Conclusion

- *RQ1: Can LLMs use deductive reasoning capabilities to understand a new task that shares the same dataset they have been trained on to solve another task?* **It is unclear to what extent LLMs can use reasoning, but they can apply pattern-matching reasoning to navigate a new task using a familiar dataset**
- *RQ2: How effectively can LLMs use intrinsic knowledge to solve a new task?* **LLMs can leverage previously amassed knowledge to successfully perform well on an unseen task**
- Work opens up new avenues in experimental settings to test LLM reasoning and knowledge probing

Thank You

Questions