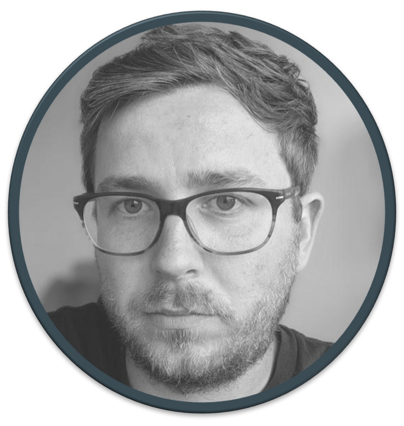
KAMEL: Knowledge Analysis with Multitoken Entities in Language models

Jan-Christoph Kalo & Leandra Fichtel

Jan-Christoph Kalo

Questions: Look for this person!



j.c.kalo@vu.nl

JanCKalo

Abstract

Large language models (LMs) have been shown to capture large amounts of relational knowledge.
They can be simply probed for this factual knowledge by using cloze-style prompts. We show that, the most frequently used dataset for knowledge probing, LAMA, has several drawbacks for analyzing the knowledge capturing the behavior of LMs.
We present a novel Wikidata-based benchmark dataset, **KAMEL**, for probing relational knowledge in LMs. It covers a **broader range of knowledge**, probes for **single-, and multi-token entities**, and contains facts with **literal values**. Furthermore, the evaluation procedure is more accurate, since the dataset contains **alternative entity labels** and deals with **higher-cardinality relations**.
We show that indeed novel models perform very well on LAMA, achieving a promising F1-score of 52.90%, while only achieving **17.62% on KAMEL**. Our analysis shows that even large language models are far from being able to memorize all varieties of relational knowledge that is usually stored knowledge graphs.

1 Introduction

- We propose a new dataset for exploring relational world knowledge in pre-trained language models
- We overcome shortcomings of the existing LAMA dataset
- Several causal language models are evaluated in a few-shot question answering setting

Prompt

Few-shot Examples

What languages does Barack Obama speak?
English, Indonesian
What languages does Chimamanda Ngozi Adichie speak?
English, Igbo, Nigerian, Pidgin

Question

(Natalie Portman, P1412, ?)
What languages does Natalie Portman speak?

Answer

English, Spanish, French, Italian

Precision 75%
Recall 60%

Gold Answer

English, Spanish, Hebrew, Japanese, French
(+ alternative labels for each answer)

2 Dataset Creation

Extraction:
- Distant supervision between Wikipedia and Wikidata filtered with a textual entailment framework to find facts that are mentioned in Wikipedia
- The result are 9,872,196 distinct triples from 1493 different Wikidata relations

Filtering:
- Remove facts (1) with literals, (2) about meta information, (3) with qualifiers, (4) too generic, (5) overlapping subject and object labels, (6) with at most 10 answers

Sampling:
- Random sampling 1000 training triples, 200 validation triples, and 200 test triples per relation for the remaining relations

3 The dataset: KAMEL

- Random Wikidata facts from 234 different relations
- All facts are mentioned in Wikipedia and therefore are part of most pre-training corpora
- Entity names have multiple tokens and therefore are often fewer known entities
- Entities have multiple aliases to guarantee a more realistic evaluation
- Queries have between 1 and 10 answers
- We probe for entities and number literals

	KAMEL	LAMA
Number of Queries	46800	31479
Number of Relations	234	41
Avg. Number of Tokens	4.86	1
Avg. Number of Labels	3.19	1
Queries with Multiple Results	4296	1035
Literals	✅	❌

WANT MORE INFO?

Download the dataset

4 Experiments

Evaluation Strategy:
- Closed-book question answering with few-shot learning and 234 manually created question templates
- Macro-averaged precision, recall, F1 scores
- Evaluation of, OPT-1.3b, GPT2-xl, GPT-J-6b, OPT-6.7b , and OPT-13b

Best and worst performing relations

Label	F1
animal breed	93.00%
continent	91.58%
languages spoken	56.41%
country	55.12%

Label	F1
shares border with	0.00%
date of death	0.00%
student of	0.00%
date of birth	0.00%

Results:
- The top performing model OPT-13b only achieves 17.62% F1-score on KAMEL, while it achieves 52.90% on LAMA
- Queries with smaller answer ranges are naturally easier to answer and achieve better performance
- Multiple aliases increase the F1 score by around 1.5%
- Queries with few gold answers can be answered easier
- Queries with numerical answers can hardly be answered correctly

Results for OPT models

	1-shot			5-shot			10-shot		
Model	P	R	F1	P	R	F1	P	R	F1
OPT-1.3b	7.02%	6.91%	6.97%	10.87%	10.61%	10.74%	11.50%	11.18%	11.34%
OPT-6.7b	10.19%	10.09%	10.14%	15.65%	15.20%	15.42%	16.67%	16.24%	16.45%
OPT-13b	10.96%	10.88%	**10.92%**	16.42%	16.22%	**16.32%**	17.76%	17.48%	**17.62%**

Conclusion

- Larger language models perform better on KAMEL than small models
- Geographic relations are often easier and can achieve high F1-scores
- Knowledge about popular entities is significantly better memorized
- Memorizing rnumbers is much more difficult than string labels
- KAMEL provides a more ealistic evaluation dataset for relational knowledge in language models
- Even large pre-trained language models cannot serve as knowledge graphs

Learning and Reasoning

ifis
Institut für Informationssysteme
Technische Universität Braunschweig