Preface: LM-KBC Challenge 2024

Jan-Christoph Kalo[1], Tuan-Phong Nguyen[2], Simon Razniewski[3] and Bohui Zhang[4]

[1]*VU Amsterdam*
[2]*MPI for Informatics*
[3]*ScaDS.AI & TU Dresden*
[4]*King's College London*

Abstract

Pretrained language models (LMs) have advanced a range of semantic tasks and have also shown promise for knowledge extraction from the models itself. Although several works have explored this ability in a setting called probing or prompting, the viability of *knowledge base construction* from LMs remains underexplored. In the 3rd edition of this challenge, participants were asked to build actual disambiguated knowledge bases from LMs, for given subjects and relations. In crucial difference to existing probing benchmarks like LAMA [1], we make no simplifying assumptions on relation cardinalities, i.e., a subject-entity can stand in relation with zero, one, or many object-entities. Furthermore, submissions need to go beyond just ranking predicted surface strings and materialize disambiguated entities in the output, which will be evaluated using established KB metrics of precision and recall. The challenge has a single track for LMs with a parameter level under 10-billion to fit into low computational requirements. The challenge received 8 submissions, of which 5 submitted a paper, and 4 were accepted for presentation. The challenge was collocated with a workshop on related topics, allowing to host extended discussions, related papers, and invited talks.

1. Introduction

Background. Large-scale LMs such as BERT [2], T5 [3], and ChatGPT [4] are optimized to predict masked-out textual inputs or perform sentence completion and have notably advanced performances on a range of downstream NLP tasks like question answering, information retrieval, machine translation and so on. Recently, LMs also gained attention for their purported ability to yield structured pieces of knowledge directly from their parameters. This is promising as current knowledge bases (KBs) such as Wikidata [5] and ConceptNet [6] are part of the backbone of the Semantic Web ecosystem, yet are inherently incomplete. While constructing a KB, major challenges include relations being optional (e.g., `place-of-death`, `cause-of-death`, or `parent-organization`) and presence of multiple correct object-entities per subject-relation pair (e.g., `shares-border`, `employer`, or `speaks-language`). Additionally, KBs need materialization for scrutability and consistent downstream usage.

Previous approaches to KB construction (KBC) use unstructured text [7, 8], crowdsourcing [9], or semi-structured resources [10, 11, 12]. In the seminal LAMA paper [1], Petroni et al. showed that LMs could highly rank correct object tokens when given an input prompt specifying the

KBC-LM'24: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC 2024
j.c.kalo@uva.nl (J. Kalo); tuanphong@mpi-inf.mpg.de (T. Nguyen); simon.razniewski@tu-dresden.de (S. Razniewski)

subject-entity and relation. Despite much follow-up work reporting further advancements [13, 14], and also criticism [15, 16, 17], the prospect of using LMs for knowledge base construction remains underexplored. The LAMA benchmark, and its variants, are not suited to investigate actual KB construction since they (i) evaluate on randomly sampled subject-object pairs, thus missing out on assessing per-subject recall, and on deciding whether a subject has objects at all, (ii) focus on single word object-entities due to the limitation of single masked token prediction specification of the underlying LM, and (iii) only evaluate a model's ranking abilities, but do not force it to make deliberate accept/reject decisions.

Relevance. Automated KB construction is a long-standing core topic of the Semantic Web community, spanning decades of research on knowledge extraction, knowledge consolidation, schema matching, and similar. Although pre-trained LMs emerged in the NLP community, their potential is also recently analyzed over diverse Semantic Web tasks, such as named entity recognition and linking, relation extraction, and knowledge reasoning. The focus of research around extracting factual knowledge from LMs is still on the NLP side (e.g., LAMA stands for "LAnguage Model Analysis"), and evaluations are limited to simplified settings. We aim to seize this opportunity for the Semantic Web and host this challenge to explore the benefits of using LMs for actual KB construction.

In the 3rd edition of our challenge, we invited participants to present solutions to make use of LMs for actual **KB construction** without prior information on the cardinality of relations, i.e., for a given subject-relation pair, the details on the total count of possible object-entities are absent. We require participants to submit a LM-based system that takes an input consisting of a subject-entity and relation, uses LM(s) depending on the choice of the track, generates disambiguated subject-relation-object triples, and makes actual accept/reject decisions for each generated output triple. Finally, we evaluated the resulting KBs using established precision and recall metrics.

2. Task Description

Given an input tuple of a subject-entity s and a relation r, the task is to generate all correct object-entities $[o_1, o_2, \ldots, o_k]$, by using language model probing. For example, on a sample tuple like (Greece, `shares-border`), one could probe the BERT model using a prompt template like "Greece shares a border with [MASK]", and produce country-type output surface strings as Turkey, Bulgaria, Albania, among others.

In the LM-KBC Challenge, participants are required to utilize LM-based system and produce actual disambiguated entities using Wikidata identifiers for all valid object entities (e.g., Q43 for Turkey), not just the surface label ("Turkey"). The final submitted system must be self-contained and should not invoke external calls, e.g., to web search engines. We will release a baseline disambiguation method for performance comparison.

Allowed Models In 2024, the LM-KBC Challenge places emphasis on both efficiency and performance. We are introducing a 10-billion parameter limit. This limit offers a good compromise between allowing reasonably performant models (e.g., Llama-7B [18], Mistral-7B [19]), and ensuring that participants can run models still locally, and are not overly concerned with

being overpowered by teams that can offer subscriptions to expensive models like GPT-4.

Evaluation For each test instance, predictions are evaluated by calculating precision and recall against ground-truth values. The final macro-averaged F1-score is used to rank the participating systems.

3. Dataset Construction

Compared with the previous edition [20, 21], the dataset was reduced to 5 more challenging relations with three distinctive characteristics, thereby enabling approaches that are more targeted to specific problems. The intended groups were:

1. Relations with many empty objects (e.g., the place of death of a person, the Nobel Prize Winner(s) graduated at the university, the stock exchange a company trades at, etc.);
2. Relations with many ambiguous objects (e.g., only people born in Washington, Springfield and Franklin, three of the most ambiguous US locations, or people that served in the 1st, 2nd, or 3rd army (existed in many countries and times), etc.);
3. Standard relations from the previous edition.

For each relation, up to 1000 subjects are provided for each of the train, validation, and withheld for challenge evaluation. The relations are hand-picked to ensure diversity, and the subject entities are of different types, e.g., person, country, organization, etc. The subject entity and their corresponding object entities are automatically sampled from Wikidata, with the following requirements:

1. Balance of object list length, i.e., we over-sampled longer lists to avoid the scenario where 1-object type example dominates. The same sampling strategy is applied for train, validation, and challenge evaluation sets, to maintain a uniform underlying distribution.
2. Balance of popular and long-tail subject entities, i.e., we use proxies like #total-statements or web hits to obtain roughly 50% popular, 50% long-tail subject entities per relation.
3. Balance of single and multi-token object entities.

For ease of entry, we will release the code for a baseline method that uses one textual probing pattern per relation, retains all assertions with greater than 50% relative likelihood, and also disambiguates the predictions. The participating systems will submit their output on the CodaLab [1] platform to get the final scores on the withheld test data.

4. Differences from Previous Editions

Major differences from previous editions of the LM-KBC challenge are:

1. **Less relations of more diverse types:** The 2nd edition might have overloaded participants with 20 relations. This time, we use a smaller set, divided into topical categories, to ensure participants can better tailor their efforts.

[1] https://codalab.org/

2. **Single parameter-bounded track:** We abolish the open track, which came down to a competition for chasing larger and thus more powerful models, and instead introduce a fixed limit of 10 billion parameters.
3. **Higher data quality:** With the reduction in the number of relations, we plan to spend more manual effort on ensuring that the dataset is of highest quality.

5. Final Ranking

The following table lists the final ranking of participating systems. Systems without a accompanying description were not ranked, since we could not verify whether they relied on LLM knowledge. All paper submissions will receive reviews shortly.

Paper Submission	User on CodaLab	Results in paper	Results on leaderboard	Rank	Authors
3	davidebara	0.91-0.94	0.9224	1	Davide Mario Ricardo Bara
4	marcelomachado	0.9083	0.9083	2	Marcelo de Oliveira Costa Machado, João Marcello Bessa Rodrigues, Guilherme Lima, Viviane Torres da Silva
6	Thin	0.698	0.6977	3	Thin Prabhong, Natthawut Kertkeidkachorn, Areerat Trongratsameethong
5	NadeenFathallah	0.653	0.6529	4	Arunav Das, Nadeen Fathallah, Nicole Obretincheva
2	hannaabiakl-dsti	0.6872	0.6872	-	Hanna Abi Akl
-	Borista	-	0.9131	-	-
-	Rajaa	-	0.5662	-	-
-	aunsiels	-	0.5076	-	-

Table 1
Leaderboard Results and Author Contributions.

References

[1] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases?, in: EMNLP, 2019.
[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: ACL, 2019.
[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, JMLR (2020).
[4] OpenAI, chatgpt, https://openai.com/blog/chatgpt/, 2022.
[5] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Communications of ACM (2014).

[6] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge., in: AAAI, 2017.
[7] N. Nakashole, M. Theobald, G. Weikum, Scalable knowledge harvesting with high precision and high recall, in: WSDM, 2011.
[8] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, W. Zhang, Knowledge vault: A web-scale approach to probabilistic knowledge fusion, in: ACM SIGKDD, 2014.
[9] A. Kobren, T. Logan, S. Sampangi, A. McCallum, Domain specific knowledge base construction via crowdsourcing, in: Neural Information Processing Systems Workshop on Automated Knowledge Base Construction AKBC, Montreal, Canada, 2014.
[10] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: A collaboratively created graph database for structuring human knowledge, in: ACM SIGMOD, 2008.
[11] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: ISWC, 2007.
[12] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, G. Weikum, Yago2: Exploring and querying world knowledge in time, space, context, and many languages, in: WWW, 2011.
[13] K. Guu, K. Lee, Z. Tung, P. Pasupat, M. Chang, Retrieval augmented language model pre-training, in: ICML, 2020.
[14] A. Roberts, C. Raffel, N. Shazeer, How much knowledge can you pack into the parameters of a language model?, in: EMNLP, 2020.
[15] R. T. McCoy, E. Pavlick, T. Linzen, Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, in: ACL, 2019.
[16] N. Kassner, H. Schütze, Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly, in: ACL, 2020.
[17] B. Cao, H. Lin, X. Han, L. Sun, L. Yan, M. Liao, T. Xue, J. Xu, Knowledgeable or educated guess? revisiting language models as knowledge bases, in: ACL, 2021.
[18] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[19] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[20] J.-C. Kalo, S. Singhania, S. Razniewski, J. Z. Pan, et al., LM-KBC 2023: 2nd challenge on knowledge base construction from pre-trained language models, in: CEUR-WS, volume 3577, 2023.
[21] S. Singhania, T.-P. Nguyen, S. Razniewski, LM-KBC: Knowledge base construction from pre-trained language models, the semantic web challenge on knowledge base construction from pre-trained language models, CEUR-WS (2022).