KLoB: a Benchmark for Assessing Knowledge Locating Methods in Language Models

Anonymous ACL submission

Abstract

Recently, Locate-Then-Edit paradigm has emerged as one of the main approaches in changing factual knowledge stored in the Language models. However, there is a lack of research on whether present locating methods can pinpoint the exact parameters embedding the desired knowledge. Moreover, although many researchers have questioned the validity of locality hypothesis of factual knowledge, no method is provided to test the a hypothesis for more in-depth discussion and research. Therefore, we introduce KLoB, a benchmark examining three essential properties that a reliable knowledge locating method should satisfy. KLoB can serve as a benchmark for evaluating existing locating methods in language models, and can contributes a method to reassessing the validity of locality hypothesis of factual knowledge. KLoB is publicly available at an anonymous GitHub: `https://github.com/anon6662/KLoB`.

1 Introduction

Language models have exhibited a significant capability to store factual knowledge (Roberts et al., 2020). Yet, as language models scale larger, the need to uphold the correctness and contemporaneity of stored knowledge becomes increasingly critical (Sinitsin et al., 2020), thus sparking a new area of research: knowledge editing. Among current knowledge editing techniques, **Locate-Then-Edit** paradigm (Dai et al., 2022; Meng et al., 2022a,b; Li et al., 2023) has emerged as one of the main approaches and garnered significant attention (Yao et al., 2023).

As depicted in Figure 1, by first locating parameters associated with specific knowledge and modifying while keeping the remaining parameters unchanged, Locate-Then-Edit methods can facilitate alterations to the model with very low cost (Yao et al., 2023). However, there is currently no method

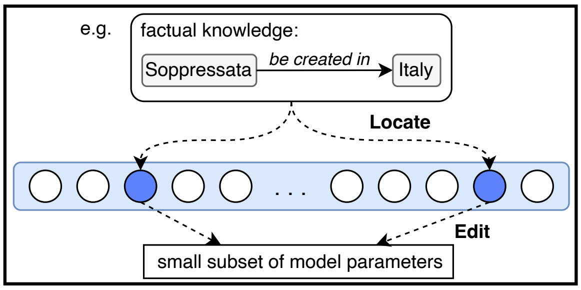

Figure 1: Illustration of Locate-Then-Edit method.

for evaluating the locating results. It's still ambiguous whether current locating methods can pinpoint the exact parameters embedding the desired knowledge. Moreover, the **locality hypothesis of factual knowledge**, which posits that factual knowledge is predominantly embedded within a small subset of parameters, has encountered a degree of skepticism and warrants further investigation. Yet, there's a noticeable absence of established methods to study and validate this concern.

Therefore, we introduce KLoB (**K**nowledge **Lo**cating **B**enchmark), a novel benchmark for evaluating locating methods in language models. KLoB delineates three essential criteria that a reliable knowledge locating method should satisfy and then evaluates the efficacy of locating methods by examining these criteria. The delineated criteria are as follows:

- **Consistency**: Locating results should remain consistent across different expressions of the same factual knowledge.
- **Relevance**: Locating results for related factual knowledge should exhibit higher similarity than those for unrelated knowledge.
- **Unbiasedness**: Parameter scores should be more uniform for inputs lacking explicit factual knowledge than those for inputs with explicit knowledge.

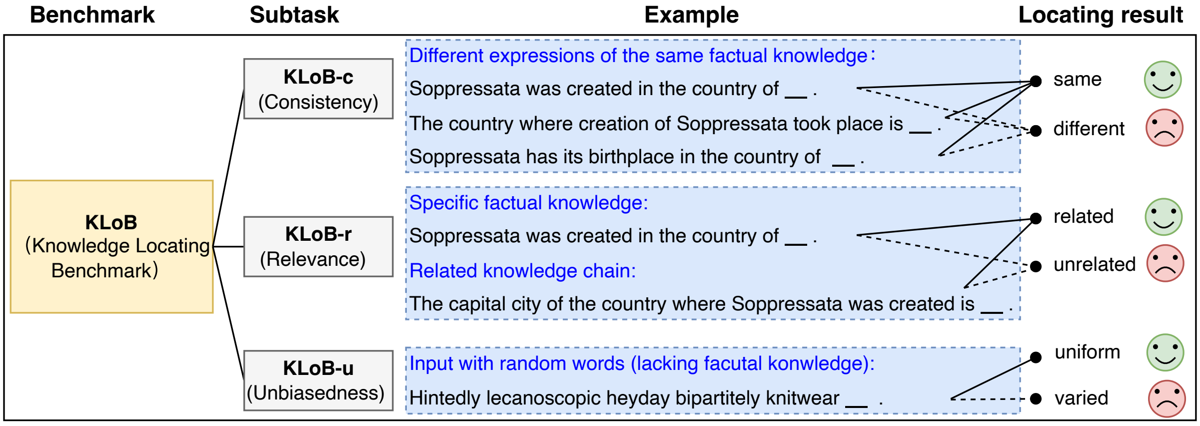

Figure 2: Examples of three subtasks in KLoB. KLoB-c: Each example comprises multiple expressions of the same factual knowledge; KLoB-r: Each example includes a sentence with specific factual knowledge and another with a related knowledge chain; KLoB-u: Each example features a sentence composed of random words.

As illustrated in Figure 2, the KLoB benchmark comprises three subtasks, each examining one of the aforementioned criteria. As the first benchmark for evaluating knowledge locating methods in language models, KLoB can play a crucial role in facilitating a comprehensive evaluation of whether current locating methods can accurately pinpoint model parameters associated with specific factual knowledge. Furthermore, KLoB can also contribute a method to study and reassess the validity of the locality hypothesis of factual knowledge.

2 Preliminary

2.1 Knowledge Editing

The factual knowledge embedded within a language model can be incorrect or become outdated over time. Knowledge editing aims to hold the correctness and contemporaneity of factual knowledge, without incurring excessive retraining costs. Current knowledge editing methods can be mainly categorized into three ways (Yao et al., 2023):

- Locate-Then-Edit: Locating and modifying model parameters that are associated with specific knowledge (Dai et al., 2022; Meng et al., 2022a,b; Li et al., 2023).
- Meta-learning: Utilizing a hyper network to learn how to change parameters of a language model (De Cao et al., 2021; Mitchell et al., 2021).
- Memory Model: Storing edits in an additional model while keeping the original model frozen (Mitchell et al., 2022; Huang et al., 2023b; Dong et al., 2022).

2.2 Locate-Then-Edit Method

Based on the locality hypothesis of factual knowledge, which posits that knowledge is primarily embedded within a subset of the model's parameters, Locate-Then-Edit methods operate in a pipeline manner: first locate a small subset of model parameters, which are associated with specific knowledge, and then modify those parameters. Because Locate-Then-Edit methods only change the located parameters while keeping the rest unchanged, they can effectively modify the model in a more targeted manner and have garnered significant attention.

3 KLoB: Knowledge Locating Benchmark

3.1 Design Philosophy

KLoB is designed for evaluating whether parameters selected by locating methods embed the desired knowledge. The underlying philosophy in constructing KLoB is to examine whether the locating results possess the desired properties. Assuming the validity of the locality hypothesis of factual knowledge, we define three criteria that a reliable knowledge locating method should satisfy: Consistency, Relevance, and Unbiasedness.

3.1.1 Consistency

Since the goal of locating methods is selecting parameters associated with specific factual knowledge, the locating result should be associated solely with the targeted knowledge, and should not be affected by other factors such as syntactic structure or synonym substitution. Therefore, we propose consistency as a criterion: **for the same factual**

knowledge, locating results should remain invariant despite variations in expression.

3.1.2 Relevance

Huang et al. (2023a) introduced the concept of multi-hop knowledge editing, where an input may be linked to a chain of interconnected factual knowledge. Locating methods should be able to recognize the correlation between the specific knowledge and the knowledge chain that includes it. Therefore, we propose relevance as a criterion: **the locating results for specific knowledge and its related knowledge chain should exhibit greater similarity compared to those for unrelated knowledge**.

3.1.3 Unbiasedness

Knowledge locating methods score and rank parameters based on their association with the targeted knowledge. Since the differences in parameter scores arise from the knowledge present in the input, for inputs that do not align with any factual knowledge, the parameter scores should be more uniform compared to those aligned with specific factual knowledge. Therefore, we propose unbiasedness as a criterion: **compared to inputs explicitly pointing to factual knowledge, parameter scores for inputs devoid of factual knowledge should be more uniform**.

3.2 Data Format

As depicted in Figure 2, KLoB consists of three subtasks, each examining one of the aforementioned criteria:

- KLoB-c (consistency): In this subtask, each example comprises three sentences of the same factual knowledge. As shown in Figure 2, both *'Soppressata was created in the country of _'* and *'The country where the creation of Soppressata took place is _'* include the factual knowledge [Soppressata $\xrightarrow{\text{created in}}$ Italy].

- KLoB-r (relevance): Here, each example comprises a sentence that includes specific factual knowledge and another one associated with its related knowledge chain. As depicted in Figure 2, the sentences *'Soppressata was created in the country of _'* and *'The capital city of the country where Soppressata was created is _'* correspond to the factual knowledge [Soppressata $\xrightarrow{\text{created in}}$ Italy] and its related knowledge chain [Soppressata, $\xrightarrow{\text{created in}}$ Italy, $\xrightarrow{\text{capital}}$ Rome)] respectively.

- KLoB-u (unbiasedness): Each example in this subtask features a sentence composed of random words, which is considered devoid of factual knowledge.

subtask	relations	avg length	examples
KLoB-c	32	8.75	13675
KLoB-r	35	20.4	9548
KLoB-u	/	10.1	25470

Table 1: Data statistics of KLoB benchmark.

As depicted in Figure 2, the answer entity in factual knowledge is positioned at the end of the sentence in KLoB. This design distinguishes KLoB from previous benchmarks (Elazar et al., 2021; Huang et al., 2023a) that utilize sentences with ['MASK'] or question sentences as input text. Consequently, examples in KLoB are compatible with both auto-regressive models, such as GPT (Radford et al., 2018, 2019) and Llama (Touvron et al., 2023), and autoencoding models, such as BERT (Kenton and Toutanova, 2019) and ALBERT (Lan et al., 2019).

3.3 Data Construction

KLoB is constructed based on Wikidata(Vrandečić and Krötzsch, 2014) and MQUAKE benchmark (Zhong et al., 2023). Table 1 summarizes the statistics of KLoB benchmark.

- KLoB-c is built upon Wikidata, a knowledge base consisting of millions of factual triples. We select relationships from Wikidata and manually construct three templates for each relationship. These templates are manually written by human experts, ensuring diversity in grammatical structures and words. Table 3 in the Appendix lists all templates for constructing KLoB-c. Then, we use these templates and corresponding entities in Wikidata to generate sentences.

- KLoB-r is constructed based on MQUAKE, a dataset comprising multi-hop questions corresponding to chains of facts. We select two-hop fact chains from MQUAKE and use the first fact in chains to generate sentences containing single fact, employing the same template as

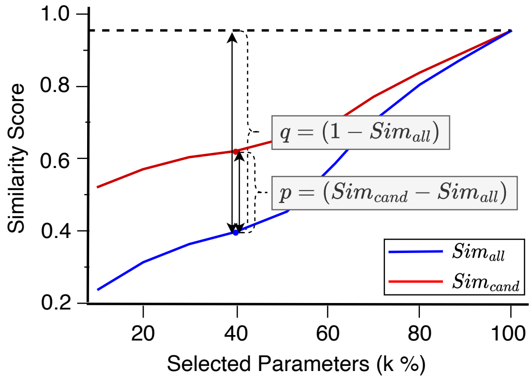

Figure 3: $RSim$: evaluation metrics for KLoB-c and KLob-r. $RSim$ is given by $RSim = max(\frac{p}{q}, 0)$.

utilized in KLoB-c. For entire fact chains, sentences are generated by rephrasing the multi-hop questions in MQUAKE.

- KLoB-u is constructed by replacing the words in the examples from KLoB-c and KLoB-r with random words and punctuation. The replacement words are sourced from the English word list of the NLTK (Natural Language Toolkit) library (Bird et al., 2009).

We utilize Llama2-7b (Touvron et al., 2023) to filter out factual knowledge that is hard to recall. We query Llama2-7b using an in-context learning approach with 8 demonstration examples and retain only the factual knowledge for which the model can correctly predict the answers.

3.4 Evaluation Metrics

In this work, we introduce the Relative Similarity ($RSim$) metric for KLoB-c and KLoB-r, assessing the extent to which the similarity within certain locating results exceeds the similarity between these and other locating results. Similarly, we introduce the Relative Standard Deviation (RSD) metric for KLoB-u, assessing how much more uniform the locating results are for sentences devoid of factual knowledge (KLoB-u) compared to those with explicit factual knowledge (KLoB-c and KLoB-r).

- **Relative Similarity**: The $RSim$ metric operates by considering only the selected parameters in locating results. The first step in the $RSim$ metric is to calculate the intra-similarity of the candidate locating results (e.g., locating results for sentences in one KLoB-c example), which is denoted as Sim_{cand}. Then, $RSim$ calculates the similarity between the candidate locating results and locating results of all samples in the subtask, which is denoted as Sim_{all}. As depicted in Figure 3, the $RSim$ metric is is calculated using the formula: $RSim = max(\frac{Sim_{cand} - Sim_{all}}{1 - Sim_{all}}, 0)$. Given that the upper bound of similarity value Sim is 1, this formula quantifies the extent to which Sim_{cand} is closer to the upper bound compared to Sim_{all}. If the candidate locating results are all identical, meaning that Sim_{cand} equals 1, then $RSim$ is equal to 1. If Sim_{cand} is less than or equal to Sim_{all}, then $RSim$ is equal to 0. The detailed formula for calculating $RSim$ is elaborated in Appendix Section A.

- **Relative Standard Deviation**: Unlike $RSim$, the RSD metric operates by considering all model parameters. RSD utilizes the standard deviation to quantify the variability of parameter scores in one locating result, denoted as SD. RSD first calculates the average SD of sentences in KLoB-c and KLoB-r (sentences with factual knowledge), denoted as $SD_{factual}$. Then, it calculates the average SD of sentences in KLoB-u (sentences without factual knowledge), denoted as $SD_{nonfactual}$. RSD is given by $RSD = \max\left(1 - \frac{SD_{nonfactual}}{SD_{factual}}, 0\right)$. RSD measures the relationship between of $SD_{nonfactual}$ and $SD_{factual}$ by calculating the proportion between them. If the parameter scores for sentences without knowledge are all identical, $SD_{nonfactual}$ equals 0, which means RSD equals 1. Conversely, if the parameter scores for sentences without knowledge are more variable than those with factual knowledge, meaning that $SD_{nonfactual} > SD_{factual}$, then RSD equals 0.

4 Conlusion

In this work, we introduce KLoB, the first benchmark for evaluating locating methods in language models. KLoB delineates three essential properties a reliable knowledge locating method should satisfy: Consistency, Relevance, and Unbiasedness. Thus, the evaluation of locating methods can be conducted by examining these properties. We hope KLoB can serve as a benchmark for evaluating existing locating methods, and contributes a quantitative analysis methods for reassessing the validity of locality hypothesis of factual knowledge.

5 Limitations

There are two primary limitations in this study that remain unexplored: **(i)** KLoB is only applicable for comparing knowledge locating methods with the same parameter granularity. For instance, a method with neuron-level granularity versus another at a model layer level cannot be effectively compared using KLoB. **(ii)** While this work introduces KLoB for evaluating the effectiveness of knowledge locating methods, it did not conduct experiments based on existing methods, which will be the primary focus of our subsequent research efforts.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023a. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023b. Transformer-patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2023. Pmet: Precise model editing in a transformer. *arXiv preprint arXiv:2308.08742*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. *arXiv preprint arXiv:2004.00345*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

A Detailed Description for calculating RSim

Suppose we have two locating results x_i and x_j, each contains $k\%$ of the model parameters. The similarity between these two locating results is given by $Sim = overlap_{ij}/(N * k\%)$, where $overlap_{ij}$ denotes the number of overlapping parameters between x_i and x_j, N is the total number of model parameters. For each example in KLoB-c, comprising three sentences, the similarity Sim_{cand} is calculated as the average of the pairwise similarities among these sentences." Similarly, Sim_{all} represents the average similarity between the selected locating result x_{cand} and all other locating results in the subtask, and is defined as $Sim_{all} = \frac{1}{M}\sum_{i=1}^{M} Sim(x_i, x_{cand})$, where M denotes the number of examples in the subtask. To streamline the calculations, we approximate Sim_{all} as $Sim(x_{all}, x_{cand})$, where x_{all} refers to a locating result derived by selecting parameters based on their average scores across the entire subtask.

B Templates of KLoB-c

Table 3 shows the relationship templates in KLoB-c, and Table 2 shows example counts for each relation.

relation	**example counts**
P103 (native)	851
P1001 (legal-term)	526
P101 (work)	198
P106 (by-profession)	292
P108 (works-for)	263
P127 (owned-uy)	330
P1303 (play)	264
P131 (located-in)	137
P136 (plays-music)	521
P1376 (capital)	140
P138 (is-name-after)	214
P1412 (communicate)	120
P159 (headquarter)	590
P17 (is-located)	783
P176 (produce)	831
P178 (develop)	702
P19 (born)	245
P276 (locate)	324
P36 (capital)	377
P37 (official-language)	667
P39 (position)	280
P407 (write)	662
P413 (play)	649
P449 (air)	567
P495 (create)	649
P26 (spouse)	266
P50 (author)	828
P112 (founded by)	181
P69 (educated at)	116
P140 (affiliated-with)	349
P175 (performer)	386
P641 (sport)	367

Table 2: Number of examples of each relation in KLoB-c

relation	templates
P103	The native language of **[X]** is **[Y]** The mother tongue of **[X]** is **[Y]** The language **[X]** speaks at hometown is **[Y]**
P1001	**[X]** is a legal term in **[Y]** **[X]** serves as one of the legal term for **[Y]** **[X]** : the legal term for **[Y]**
P101	**[X]** works in the field of **[Y]** **[X]** specializes in the field of **[Y]** The domain of activity of **[X]** is **[Y]**
P106	**[X]** works as **[Y]** **[X]** 's occupation is **[Y]** the profession of **[X]** is **[Y]**
P108	**[X]** works for the company: **[Y]** The company that employs **[X]** is **[Y]** The company providing employment to **[X]** is **[Y]**
P127	**[X]** belongs to **[Y]** the company that owns **[X]** is **[Y]** the owner company of **[X]** is **[Y]**
P1303	**[X]** plays the musical instrument known as the **[Y]** **[X]** is known for playing the **[Y]** In the hands of **[X]** , music emerges from the **[Y]**
P131	**[X]** is located in **[Y]** **[X]** can be found in **[Y]** The location of **[X]** is **[Y]**
P136	**[X]** composes in the genre of **[Y]** **[X]** engages in the performance of **[Y]** The music genre that **[X]** performs is **[Y]**
P1376	**[X]** is the capital of **[Y]** **[X]** holds the status of being the capital of **[Y]** **[X]** , the capital city of **[Y]**
P138	**[X]** is the capital of **[Y]** **[X]** holds the status of being the capital of **[Y]** **[X]** , the capital city of **[Y]**
P1412	**[X]** used to communicate in the language of **[Y]** **[X]** expressed himself through the language of **[Y]** In language **[X]** utilizes for communication is **[Y]**
P159	The headquarters of **[X]** is located in **[Y]** **[X]** , whose headquarters is in **[Y]** **[X]** has established its headquarters in **[Y]**
P17	The headquarters of **[X]** is located in **[Y]** **[X]** , whose headquarters is in **[Y]** **[X]** has established its headquarters in **[Y]**
P176	**[X]** is produced by the company: **[Y]** The company behind **[X]** Lumia 800 **[Y]** **[X]** is one among products crafted by **[Y]**
P178	**[X]** , a product manufactured by the company: **[Y]** The company that developed **[X]** is **[Y]** The company that stands behind the creation of **[X]** is **[Y]**
P19	**[X]** was born in **[Y]** The place of birth for **[X]** is **[Y]** The birth of **[X]** occurred in **[Y]**
P276	**[X]** is located in **[Y]** **[X]** can be found in **[Y]** The location of **[X]** is **[Y]**
P36	**[X]** is the capital of **[Y]** **[X]** holds the status of being the capital of **[Y]** **[X]** , the capital city of **[Y]**
P37	The official language of **[X]** is **[Y]** In terms of official language, **[X]** uses **[Y]** Under **[X]** law, the official language is recognized as **[Y]**
P39	**[X]** assumed the role of **[Y]** **[X]** holds the position of **[Y]** **[X]** served in the capacity of **[Y]**
P407	The language of **[X]** is **[Y]** The **[X]** was penned in **[Y]** **[X]** was written with the langeuage of **[Y]**
P413	**[X]** plays in the position of a **[Y]** **[X]** 's role on the team involves him serving as a **[Y]** When on the field, **[X]** is positioned as a **[Y]**
P449	**[X]** was originally aired on **[Y]** **[X]** , initially, was broadcasted on **[Y]** **[X]** was originally presented to audiences on **[Y]**
P495	**[X]** was created in the country of **[Y]** The country where creation of **[X]** took place is **[Y]** **[X]** has its birthplace in the country of **[Y]**
P26	**[X]** is married to **[Y]** The spouse of **[X]** is none other than **[Y]** In matrimony, **[X]** is bound to **[Y]**
P50	The author of **[X]** is **[Y]** **[X]** was written by **[Y]** Credited with the creation of **[X]** is **[Y]**
P112	**[X]** was founded by **[Y]** The **[X]** owes its existence to the person of **[Y]** The person behind the inception of the **[X]** is **[Y]**
P69	**[X]** was educated at a university named **[Y]** **[X]** received his education from the institution known as **[Y]** The university where **[X]** was educated is **[Y]**
P140	**[X]** is affiliated with the religion of **[Y]** **[X]** is a follower of the faith known as the **[Y]** The religion that **[X]** adheres to is **[Y]**
P175	**[X]** was performed by **[Y]** **[X]** was presented to audiences by **[Y]** In the performance of **[X]** , the artist is **[Y]**
P641	**[X]** is associated with the sport of **[Y]** The sport that **[X]** is linked to is association **[Y]** **[X]** pertains to the sport known as association **[Y]**

Table 3: Templates of KLoB-c, where [X] refers to the head entity of fact triples in Wikidata, and [Y] refers to the tail entity.