Entity Cloze By Date:
Understanding what LMs know about unseen entities

Anonymous ACL submission

Abstract

Language models (LMs) are typically trained once on a large-scale corpus and used for years without being updated. Our world, however, is dynamic, and new entities constantly arise. We propose a framework to analyze what LMs can infer about new entities that did not exist when the LMs were pretrained. We derive a dataset of entities indexed by their origination date and paired with their English Wikipedia articles, from which we can find sentences about each entity. We evaluate LMs' perplexity on masked spans within these sentences. We show that models more informed about the entities, such as those with access to a textual definition of them, achieve lower perplexity on this benchmark. Our experimental results demonstrate that making inferences about new entities remains difficult for LMs. Given its wide coverage on entity knowledge and temporal indexing, our dataset can be used to evaluate LMs and techniques designed to modify or extend their knowledge. Our automatic data collection pipeline can be easily used to continually update our benchmark.

1 Introduction

New entities arise every day: new movies, TV shows, and products are created, new events occur, and new people come into the spotlight. Whatever the capabilities of language models (LMs) to represent entity knowledge, these new entities cannot possibly be included in the language models' parametric knowledge (i.e., knowledge acquired during pretraining), as they did not exist when LMs were trained. Since this temporal mismatch between LMs and real-world knowledge affects model performance on downstream tasks (Zhang and Choi, 2021; Dhingra et al., 2021; Lazaridou et al., 2021), understanding what LMs know about real-world entities is an important task.

The existing literature provides various benchmarks to measure LMs' knowledge about entities (Petroni et al., 2019, 2021; Dhingra et al., 2021). Those benchmarks are typically formulated as cloze-style tasks covering a limited set of relations in the knowledge bases (e.g., Wikidata). The knowledge to be tested is bounded by the underlying KB relations, e.g., LAMA uses around 40 Wikidata relations and entities collected in 2017. There have been LAMA-like probing tasks that consider temporal aspects (Dhingra et al., 2021; Jang et al., 2021), but those datasets do not differentiate new and existing entities. Therefore, the current knowledge probing datasets are not suitable for our purpose: testing broad knowledge about real-world entities, and how LM's knowledge differ on entities that are unseen during pre-training for temporal mismatch and entities that could have occurred in pre-training corpus.

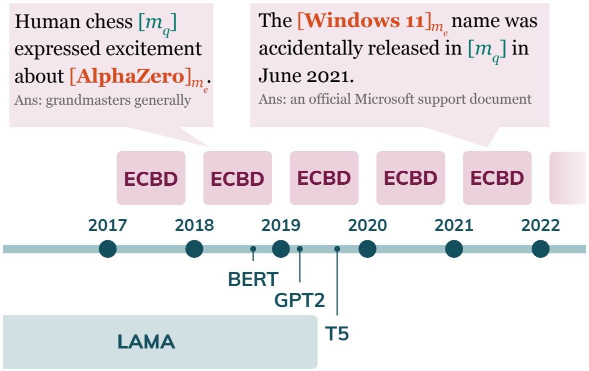
Human chess $[m_q]$ expressed excitement about [AlphaZero]$_{m_e}$.
Ans: grandmasters generally
The [Windows 11]$_{m_e}$ name was accidentally released in $[m_q]$ in June 2021.
Ans: an official Microsoft support document
ECBD
ECBD
ECBD
ECBD
ECBD
2017
2018
2019
2020
2021
2022
BERT
GPT2
T5
LAMA

Figure 1: Overview of knowledge coverage by our dataset and LAMA. Our framework collects entities from certain time period.

To fill this gap, we propose a framework to evaluate LMs' knowledge about entities classified by their origination date. We extract a set of Origination Date Indexed Entities (ODIE) based on metadata from Wikidata. We then construct cloze statements (i.e., masked sentences) from those entities' Wikipedia articles. Unlike past knowledge probing datasets, these cloze sentences test the ability of a model to make a wide range of inferences related to entities, without being resticted to a pre-defined set of KB relations. We choose masked spans near these entities that likely contain information related to the entities, which we evaluate based on the perplexity gap between the raw sentence and the sentence with the entity replaced.

We release an Entity Cloze by Date (ECBD) dataset of 40k masked sentences that contain mentions of 2.3K ODIE entities,[1] split by year covering a time period from 2017 to 2021. In our experiments, we evaluate a pre-trained language model (Raffel et al., 2020) in terms of perplexity and exact match performance measured by recall@10. We establish that injecting additional information such as a text definition can meaningfully teach the model to make better guesses about masked spans, highlighting this dataset's utility for benchmarking methods of knowledge injection.

2 Entity Cloze by Date

We aim to test language models' 1) broader entity knowledge and 2) ability to reason about completely unseen entities (i.e., unseen during pretraining). Towards these goals, we want to have the following properties in our entity cloze sentences. **(1) Date indexing.** If each cloze example is associated with an entity and indexed by the origination date of that entity, we can understand whether a model may have seen it in its pre-training corpus or not. **(2) Diverse sentences.** When going beyond KB triples, entity knowledge can take many forms: actions that an entity can take, other entities that action can effect, typical ways in which an entity is described, and more. Thus, we want include diverse sentences and masked spans that cover rich relations and various syntactic categories (e.g., POS and nonterminal categories, span length).

2.1 Task Definition

Given an input sentence s, containing an entity mention span m_e and a masked query span m_q, a language model is asked to predict the gold masked span m_y. Each entity e is paired with e_i, its origination year. See the following example:

e: RNA vaccine, e_i: 2020
s: [mRNA vaccines]$_{m_e}$ do not affect or reprogram [m_q].
m_y: DNA inside the cell

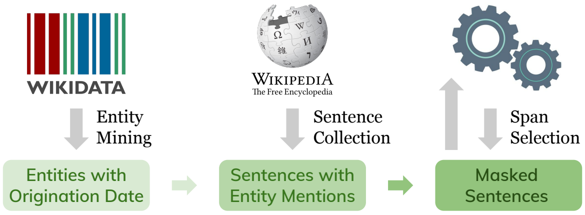

Figure 2: Overview of the data collection process.

We evaluate language models by two scores: **perplexity** on the masked span m_q and **recall@10**, i.e., a binary score indicating if model's top ten predictions contains the gold masked span m_y.

2.2 Data Collection

We divide our data collection protocol into three stages: *entity mining*, *sentence collection* and *span selection*. In this work, we use English Wikipedia (the September 1, 2021 dump) and Wikidata as our knowledge sources.

ODIE Mining We begin by gathering all entities on Wikidata that have an associated *start time*, *announcement date*, *time of discovery or invention*, *inception* date, *point in time*, or date it was *introduced on*. For such entities, we take the first of these dates to create our temporal splits, assuming that this is the earliest date the entity could have appeared in any pretraining corpus.

Entity Sentence Collection Once we obtain a list of entities, we look up their English Wikipedia articles. To enrich the candidate sentence pool and exclude trivial sentences from stub articles, we filter entities if their corresponding articles contain less than 500 words. From each article, we exclude the first paragraph of the article, to be used as an entity definition, and sample sentences from the rest of the paragraphs. We sample sentences that include the entity name or one of their Wikidata aliases. We do not accept entity mention spans located in quotes since they are often in nested named entities such as book titles.We also filter out any sentences with less than five words.

Span Selection Next, we determine spans m_q to be masked on a sentence, s; we can have multiple masked spans per sentence. All spans must be: (a) not overlapping with the entity mention span, m_e; (b) starting no more than ten words away from the mention span, to improve relatedness to the entity.

We extract two types of spans: **NP spans** are selected from any suitable noun phrases in the sentence using spaCy (Honnibal and Montani, 2017). These spans primarily represent relational knowledge about the entity, analogous to the object in a

[1] We release our datasets and software upon publication.

Origination Year	2017	2018	2019	2020	2021	Total	Example Entities
# Dev Entities	324	313	234	207	85	1,141	
# Test Entities	315	307	230	196	93	1,163	
Sports	20	18	23	13	27	19	2017 Tour de France, USL League One, Evo 2017
Media	19	20	24	23	21	21	Emily in Paris, Luigi's Mansion 3, The Midnight Gospel
Infrastructure	10	7	10	7	9	9	Gateway Arch National Park, Istanbul Airport, I-74 Bridge
Products	4	3	4	3	2	3	Apple Card, Sputnik V COVID-19 vaccine, Pixel 4
Businesses	15	11	8	6	2	10	Raytheon Technologies, Electrify America, Good Party
Organizations	15	17	12	11	10	14	NUMTOT, UK Student Climate Network
Natural Risks	3	5	4	15	11	7	Hurricane Ida, COVID-19, North Complex Fire
Other Events	9	13	12	14	7	12	Super Bowl LIV halftime show, Storm Area 51

Table 1: Origination date indexed entity (ODIE) statistics by category. The number represents % of entities with particular type among entities originated in that year.

KB triple. **Random spans** are arbitrary sequences of words sampled from the sentence. This broader set of spans may cover other types of entity knowledge (e.g., probable actions an entity can take). We uniformly sample span length between 1 and 5 and then randomly select the starting location of the span within the sentence. We only accept valid spans not overlapping with the entity mention. We extract at most 100 spans per entity to limit any one entity's contribution to the final dataset.

Constructing Popular Entity Cloze Set To compare with ODIE which covers relatively new entities originated in 2017 at the earliest, we use a set of popular entities ranked by article contributor numbers and incoming links from prior work (Onoe et al., 2021; Geva et al., 2021). We follow the same sentence collection and span selection process as ECBD to create cloze tasks on them.

Span sensitivity to entity knowledge To see if our design choices are effective, we perform a test that measures the performance drop in perplexity using T5 when we replace the entity mention with a generic reference to "the entity." We use entities from our Popular set to ensure that the LM has seen them during pre-training. If a masked span is related to the entity, the perplexity of that span should increase when the entity mention is omitted.

We see that the median perplexity of a span increases by 32.2% when the entity is removed, indicating that these spans are indeed related to the entity. Moreover, removing the distance-based criterion for span selection decreases the perplexity change to 25.9%. These results indicate that our selected spans should be correlated with the entity. This gap test was performed only for analysis and we do not use any model-based data filtering.

Dataset Statistics Table 1 shows the statistics and examples of origination date indexed entities, split by entity types. While our entity set does not comprehensively capture all entities originated in that year, it contains a diverse set of entities, ranging from events, products to organizations. One notably missing entity category is people; it is hard to pin down an origination year because of the significant gap between birth year and the year someone became prominent.

	# Sent.	# Ent.	m_q Span Len.	|Span V.|
LAMA$_{\text{TREx}}$	34k	29,488	1.0	2,017
ECBD	40k	2,304	2.9	19,542
Popular	8k	1,945	2.9	8,619

Table 2: Data statistics. |Span V.| means the vocabulary size of masked spans. Initial release of the data sample equal number of masked sentences per year (2017-21).

Table 2 reports statistics on our cloze task data and existing probe dataset (Petroni et al., 2019). While containing fewer entities, our dataset exhibits much richer vocabulary (19K vs. 2K), demonstrating diverse knowledge it covers. We split this data into dev and test sets by entities (i.e., no shared entities between dev and test). To balance our the data sizes across the groups, we sample 4k examples for each year group, yielding 40k examples in total (20k for dev and 20k for test). Earlier dates contain a larger set of entities (639 entities for 2017 compared to 178 entities for 2021) as entities are continuously updated in Wikidata. In other words, many entities originated in 2021 have not been yet added to Wikidata. We sample the same number of NP spans and random spans. Within the NP spans, 35% of them are proper noun phrases (i.e., named entities).

3 Experiments

We verify that ECBD does test entity knowledge and can be a testbed for knowledge injection techniques to teach models about unseen entities.

	POPULAR		2017		2018		2019		2020		2021	
Input Type	PERP.	R@10	PERP.	R@10	PERP.	R@10	PERP.	R@10	PERP.	R@10	PERP.	R@10
ORIGINAL	5.55	27.6	6.24	26.6	6.39	26.4	6.22	27.0	7.19	23.8	6.70	24.7
NO ENT	7.04	24.7	7.43	23.7	8.00	22.5	7.65	23.2	9.61	21.9	8.80	22.4
RANDOM DEF.	5.23	27.4	5.67	27.9	5.86	28.3	5.77	28.1	6.48	24.9	6.12	26.2
DEFINITION	**4.90**	**29.5**	**5.04**	**32.2**	**5.11**	**30.5**	**4.99**	**30.9**	**5.66**	**28.6**	**5.19**	**29.4**
Δ(ORIG. → DEF.)	0.65	1.90	1.20	5.60	1.28	4.10	1.23	3.90	1.53	4.80	1.51	4.70

Table 3: Results of T5 (pre-trained with data from 2019) on the test set, showing perplexity (↓) and recall@10 (↑).

Setup We evaluate T5-large (Raffel et al., 2020) on our dataset in the zero-shot setting where the model parameters are fixed. In addition to the original masked sentence (ORIGINAL), we feed three modified masked sentences. NO ENT replaces the entity mention span with a generic string "the entity," to test how much a generic span changes the prediction. RANDOM DEF. prepends a definition sentence of a randomly selected entity. DEFINITION prepends the first sentence of the entity's Wikipedia article to the cloze sentence.

We evaluate the T5 model on five different subsets (from 2017 to 2021) as well as a set of popular entities. Note that the entities in the 2020 and 2021 subsets are necessarily unseen to T5.

Results Table 3 reports perplexity (lower is better) and recall@10 (higher is better) on the test set. In all subsets in both metrics, we observe three consistent trends. (1) NO ENT always degrades performance compared to ORIGINAL. This result confirms that our masked spans are sensitive to the content of the entity span, although it is not conclusive proof of entity knowledge being required, as changing to "the entity" modifies other latent stylistic attributes that T5 may be sensitive to. (2) RANDOM DEF. slightly improves performance although the additional information is taken from a random entity. This could be due to the model using different positional encodings as a result of having a definition, or LMs may select information if it is useful in some cases, leading the small gains. (3) DEFINITION always boosts performance over ORIGINAL, indicating that providing more information about entities helps to retrieve information distributed over LMs's parameters.

Performance on unseen entities We investigate the performance delta between ORIGINAL and DEFINITION per origination year. We do not see clear patterns between new entities (2019 to 2021) and existing entities (2017, 2018) for T5 model we evaluate. Some entities in the 2017 and 2018 subsets might rarely occurred in pre-training corpus; therefore, definition sentences are still useful and improve performance. However, the performance delta on the popular entity set is notably smaller than others (compare: 5.55 → 4.90 for POPULAR versus 6.22 → 4.99 for 2019). This implies that LMs do contain some prior knowledge about common entities they have observed before, and can use additional information about new entities or less frequent entities. How to inject knowledge requires further investigation.

4 Use Cases

We envision this dataset as being useful for general knowledge probing, as the real-world knowledge covered by the existing benchmarks is gradually outdated. With our framework, we can easily **update** datasets using the most recent knowledge sources with a controlled manner. Since the entity knowledge in our dataset is time-indexed, this is suitable for evaluating knowledge editing approaches (Sinitsin et al., 2020; De Cao et al., 2021; Mitchell et al., 2021) and also continual knowledge learning approaches (Jang et al., 2021). Crucially, existing work studies whether these approaches can inject single facts, but not whether they can enable models to robustly support a broad range of new inferences about entities, like our dataset allows.

Although our dataset follows the widely-used cloze format, our focus is orthogonal to datasets like the Children's Book Test (Hill et al., 2016) and LAMBADA (Paperno et al., 2016), which come from fiction and do not cover real-world entities.

5 Conclusion

In this paper, we present a dataset to understand language models' broad inferences about entities across time. We collect 40k cloze-style sentences associated with a time-indexed set of entities. We also perform analysis on our data set and show that handling completely unseen entities remains challenging for the current LMs.

References

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2021. Time-Aware Language Models as Temporal Knowledge Bases.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In *International Conference on Learning Representations (ICLR)*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2021. Towards Continual Knowledge Learning of Language Models. abs/2110.03215.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the Gap: Assessing Temporal Generalization in Neural Language Models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2021. Fast Model Editing at Scale. abs/2110.11309.

Yasumasa Onoe, Michael J.Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and R. Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *ArXiv*, abs/1606.06031.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Examples

See Table 4.

B Perplexity per span type

See Table 5.

C Recall per span type

See Table 6.

Masked Sentence	Span Type	Origin Year
At 18:00 UTC on August 16, after **Grace** exited the Dominican Republic, [MASK] were lifted. Answer: “all tropical storm watches”	NP	2021
AirTags can be [MASK] the Find My app. Answer: “interacted with using”	RANDOM	2021
British tabloid “The Sun” is credited with the first headline use of ‘**Megxit**’ on [MASK] 2020. Answer: “9 January”	NP	2020
As a result, phone cases made [MASK] will also fit the **iPhone SE**. Answer: “to fit the iPhone 8”	RANDOM	2020
The **GPT-2** model has [MASK], and was trained on a dataset of 8 million web pages. Answer: “1.5 billion parameters”	NP	2019
The epicenter of the **2019 Albania earthquake** [MASK] kilometers from Tirana to the Northwest. Answer: “was about 30”	RANDOM	2019
On November 12, 2019, **Maverick City Music** released MASK, "Maverick City, Vol. 2". Answer: “their follow-up EP”	NP	2018
Austin FC are the operators of a newly-[MASK]. Answer: “built stadium at McKalla Place”	RANDOM	2018
The NFL ultimately selected Houston as [MASK] of **Super Bowl LI**. Answer: “the host city”	NP	2017
Hurricane Irma was the top Google searched term in [MASK] in 2017. Answer: “the U.S. and globally”	RANDOM	2017

Table 4: Examples selected from the 2017-2021 subsets.

	Existing Entities				New Entities					
	2017		2018		2019		2020		2021	
Input Type	NP	Rand	NP	Rand	NP	Rand	NP	Rand	NP	Rand
Original	5.86	7.33	5.81	7.51	6.11	7.29	5.92	7.63	6.23	7.31
No Ent	5.90	8.02	5.78	8.56	5.99	8.31	6.75	9.36	7.28	9.21
Random Def.	5.59	6.60	5.54	6.84	5.77	6.60	5.70	6.98	6.01	6.65
Definition	4.96	5.98	4.98	6.02	5.12	5.85	5.14	6.13	5.13	5.82

Table 5: Results of T5 model (pre-trained with data from 2019) on the dev set with perplexity (↓) per span type.

D Data Licensing

The Wikipedia text we used is licensed under CC BY-SA. Our use of Wikipedia, constructing a dataset which we will make publicly available under the same license, is consistent with the terms of the license.

E Computational Resources

All experiments were conducted using an NVIDIA Quadro RTX 8000. We only evaluate existing models on our datasets and did not do any finetuning. One evaluation experiment typically takes 15 minutes to complete. For T5 experiments, we use Hugging Face’s Transformer package (Wolf et al., 2020).

	Existing Entities				New Entities					
	2017		2018		2019		2020		2021	
Input Type	NP	RAND	NP	RAND	NP	RAND	NP	RAND	NP	RAND
ORIGINAL	30.3	20.0	31.8	20.2	29.3	22.0	30.1	19.8	29.3	21.6
NO ENT	27.2	18.8	28.1	16.7	26.2	18.1	26.8	16.7	25.9	18.2
RANDOM DEF.	31.8	20.8	32.8	19.9	29.8	21.6	31.3	20.5	29.5	21.6
DEFINITION	34.1	22.8	35.9	22.8	33.0	24.9	33.7	23.0	32.7	25.2

Table 6: Results of T5 model (pre-trained with data from 2019) on the dev set with recall@10 (↑) per span type.