Probing Knowledge in LMs

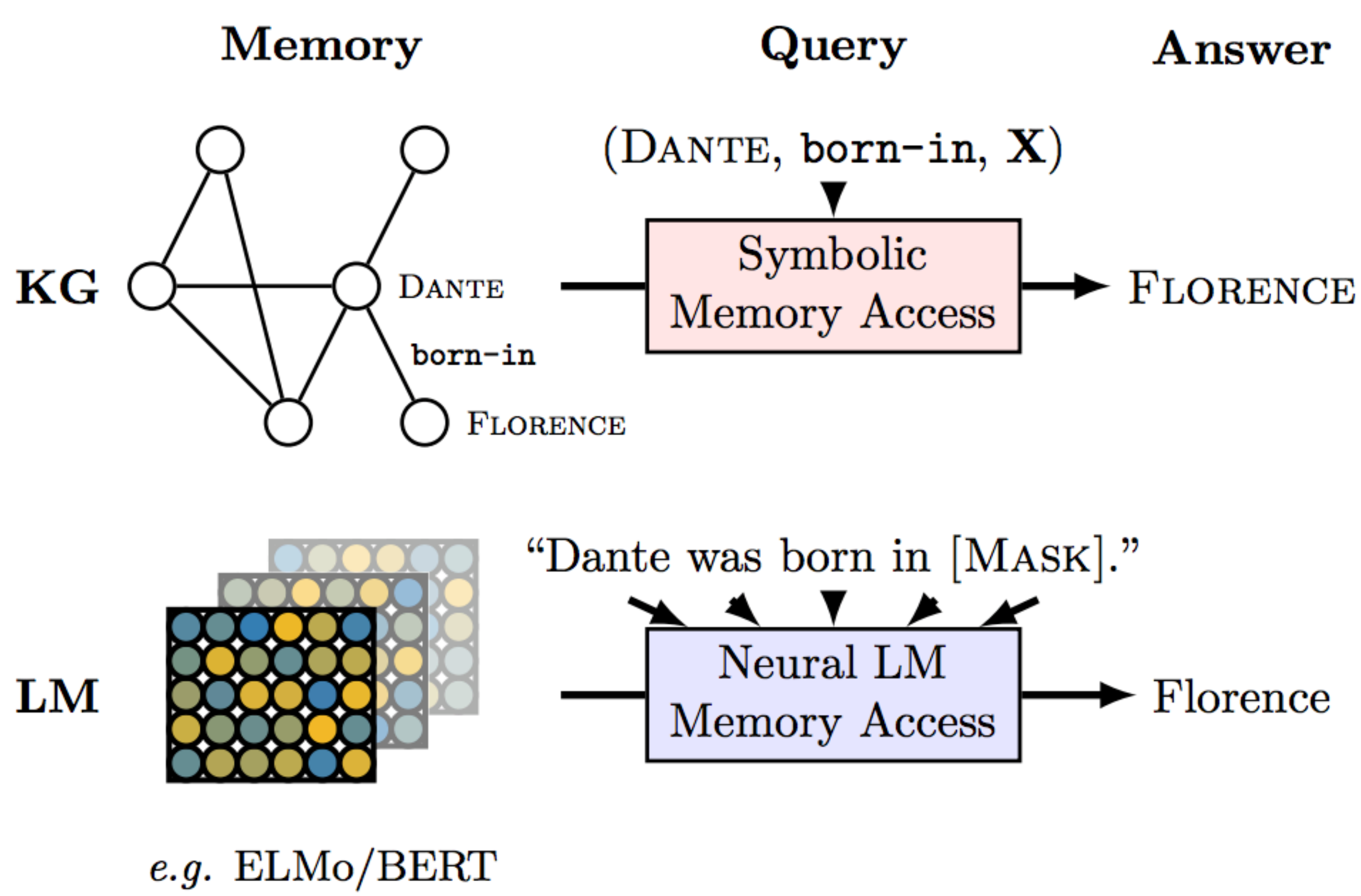
- Traditional QA/MRC models usually refer to external resources to answer questions, e.g., Wikipedia articles or KGs.

- Do LMs pre-trained on a large text corpus already capture those knowledge?

LMs as KBs?

(Petroni et al. 2019)

- Structured queries (e.g., SQL) to query KBs.
- Natural language prompts to query LMs.



LMs as KBs?
(Petroni et al. 2019)

- LAMA benchmark
  - Manual prompts for 41 relations: “[X] was born in [Y].”
  - Fill in subjects and have LMs (e.g., BERT) predict objects: “Barack Obama was born in [MASK].”
  - Accuracy: ELMo 7.1%, Transformer-XL 18.3%, BERT-base 31.1%

Prediction	Score
Barack Obama was born in **Chicago** .	5.9%
Barack Obama was born in **Philadelphia** .	3%
Barack Obama was born in **Illinois** .	1.5%
Barack Obama was born in **Detroit** .	1.4%
Barack Obama was born in **Pennsylvania** .	1.4%

https://demo.allennlp.org/masked-lm/s/barack-obama-was-born-mask/D8T2D0I0O9

How Can We Know What LMs Know? (Jiang et al. 2019)

- Query LMs with different prompts might lead to different predictions.
- Ensemble multiple mined/paraphrased prompts further increase the accuracy: 31.1% → 39.6%

Prompts

manual	DirectX *is developed by* y_{man}
mined	y_{mine} *released the* DirectX
paraphrased	DirectX *is created by* y_{para}

Top 5 predictions and log probabilities

	y_{man}		y_{mine}		y_{para}	
1	Intel	-1.06	Microsoft	-1.77	Microsoft	-2.23
2	Microsoft	-2.21	They	-2.43	Intel	-2.30
3	IBM	-2.76	It	-2.80	default	-2.96
4	Google	-3.40	Sega	-3.01	Apple	-3.44
5	Nokia	-3.58	Sony	-3.19	Google	-3.45

AutoPrompt: Automatically Generated Prompts: (Shin et al. 2020)

- Search tokens in the prompts (i.e., trigger tokens [T]) guided by gradients that maximize the probability of correct answers.

- Further increase the accuracy: 39.6% → 43.3%

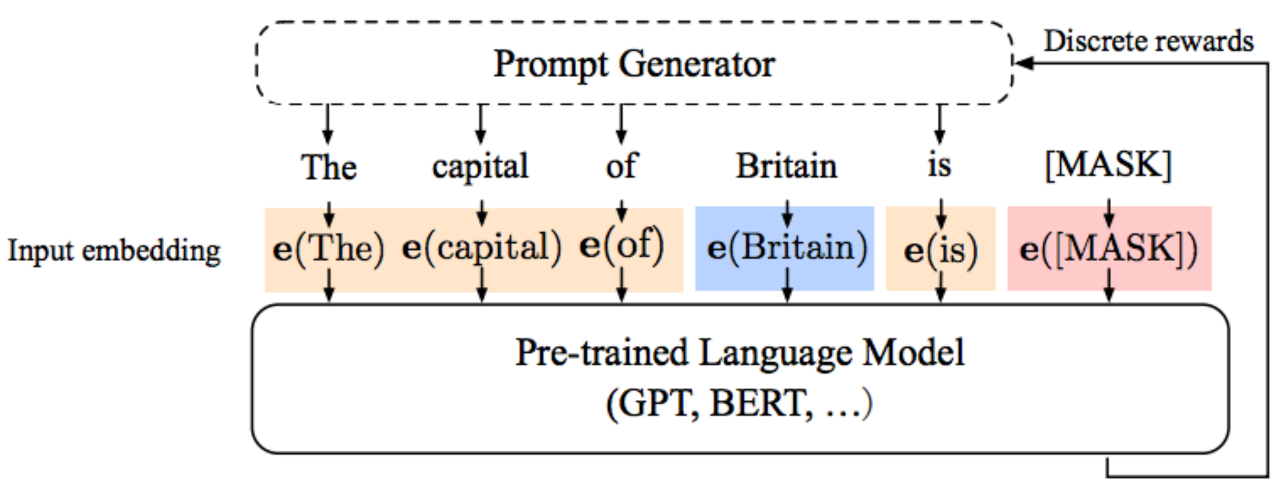
X plays Y music
{sub}[T]. . . [T][P].

Hall Overton fireplacemade antique son alto [MASK].

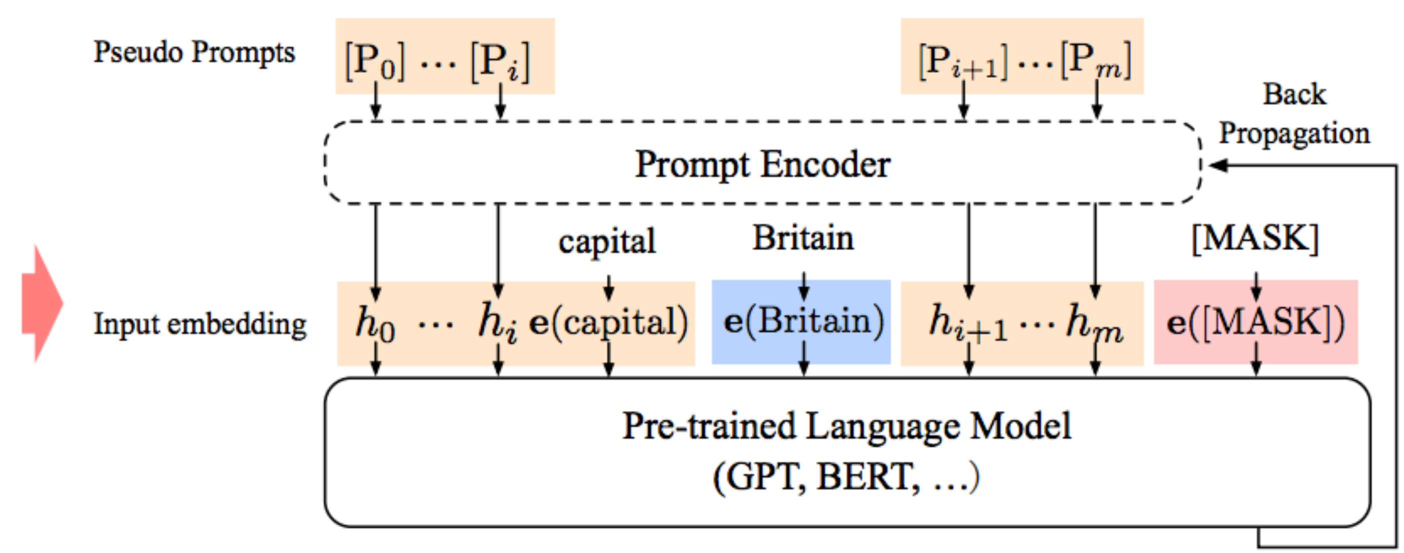
Relation	Method	Prompt	P@1
P101	Manual	[X] works in the field of [Y]	11.52
	AUTOPROMPT BERT	[X] probability earliest fame totaled studying [Y]	15.01
	AUTOPROMPT RoBERTa	[X] 1830 dissertation applying mathsucci [Y]	0.17
P103	Manual	The native language of [X] is [Y]	74.54
	AUTOPROMPT BERT	[X]PA communerug speaks proper [Y]	84.87
	AUTOPROMPT RoBERTa	[X]neau optionally fluent!?ťraditional [Y]	81.61
P106	Manual	[X] is a [Y] by profession	0.73
	AUTOPROMPT BERT	[X] supporters studied politicians musician turned [Y]	15.83
	AUTOPROMPT RoBERTa	[X] (), astronomers businessman·former [Y]	19.24
P127	Manual	[X] is owned by [Y]	36.67
	AUTOPROMPT BERT	[X] is hindwings mainline architecture within [Y]	47.01
	AUTOPROMPT RoBERTa	[X] picThom unwillingness officially governs [Y]	39.58

P-tuning: Directly Optimize Embeddings (Liu et al. 2021)

- Optimizing embeddings (continuous) is easier than searching tokens (discrete).
- Further increase the accuracy: 43.3% → 48.3%



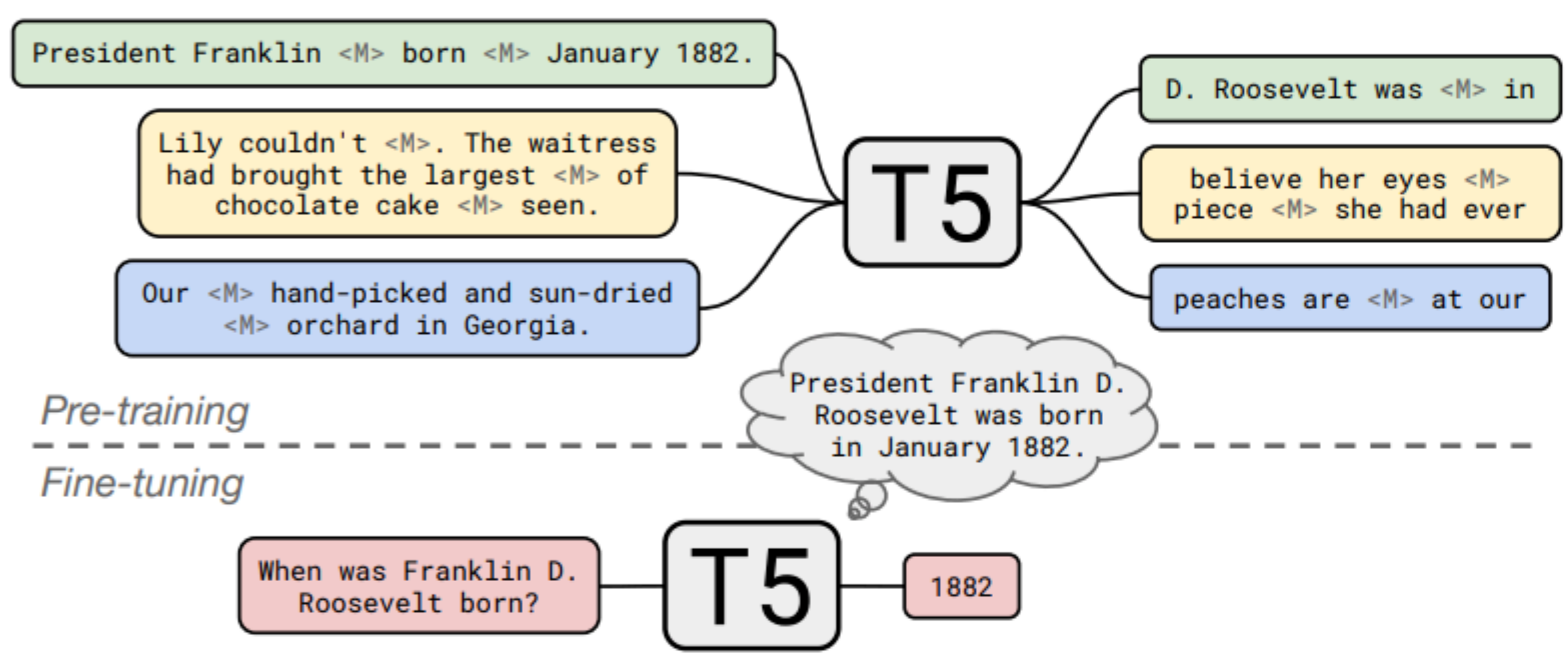
(a) Discrete Prompt Search



(b) P-tuning

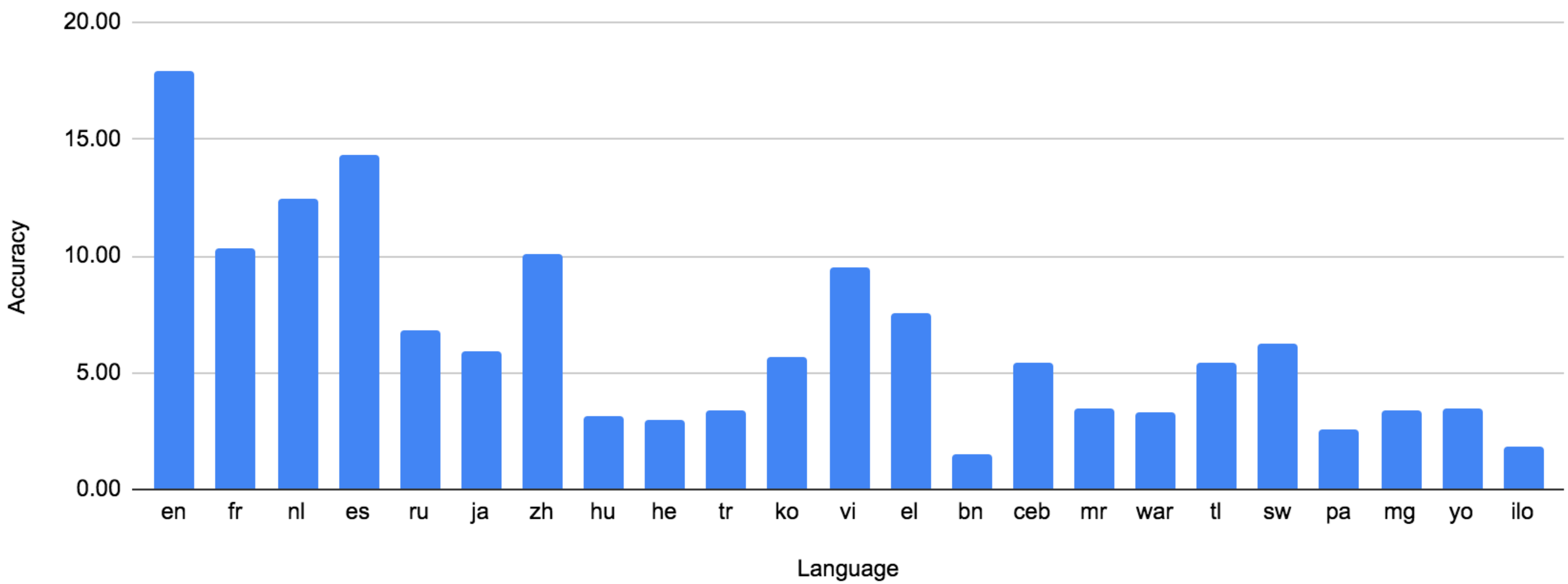
Close-book T5: Directly Fine-tune with QA Pairs (Roberts et al. 2020)

- Generate answers given questions without additional context.
- Performs even better than QA models with retrieved context (such as DrQA).



X-FACTR: Multilingual Factual Knowledge Probing (Jiang et al. 2020)

- Overall, factual knowledge in LMs is still limited, especially for low-resource languages.



Max performance of M-BERT, XLM, XLM-R

Nonparametric Models Outperform Parametric Models

- For knowledge-intensive tasks like QA, nonparametric models (w/ retrieved context) outperform parametric models (w/o context) by a large margin.

- For example, REALM (Guu et al. 2020), RAG (Lewis et al. 2020) on the NaturalQuestion datasets.

Close-book T5	34.5
REALM	40.4
RAG	44.5