Relational Knowledge and Language Models

Cardiff NLP Reading Group

Asahi Ushio
Cardiff University

Language Model Pretraining

- Large scale language model is now a huge trend
- Many architectures (seq2seq, uni/bi-directional, non-autoregressive, external knowledge, etc)
- Growing data/model size

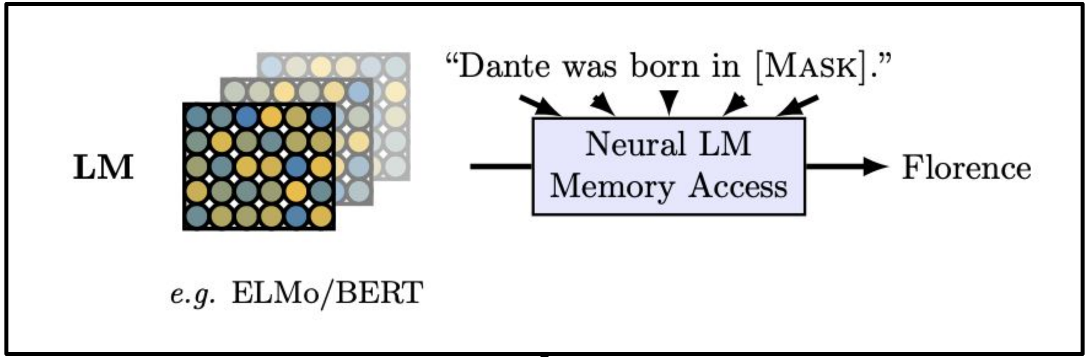
What actually language model knows?

Agenda

An overview of recent study in **“relational knowledge understanding in pretrained language models”**

- *Petroni, et al. "Language models as knowledge bases?" 2019*
- *Jiang, et al. "How can we know what language models know?" 2019*
- *Bouraoui, et al. "Inducing relational knowledge from BERT." 2019*

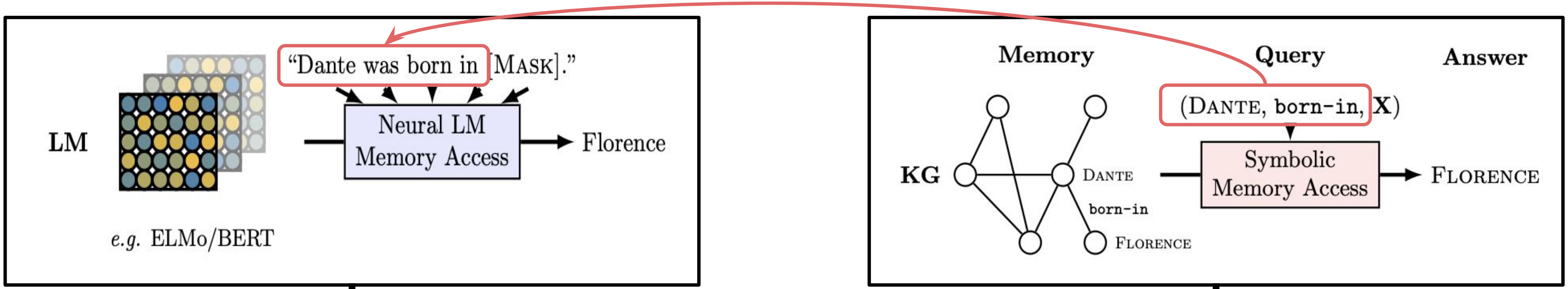
What LM knows?



LM knows the fact
"Dante was born in Florence"

[Petroni, et al. "Language models as knowledge bases?"]()

It’s a knowledge graph!



LM knows the fact
“Dante was born in Florence”

KG includes the fact
“(Dante, born-in, Florence)”

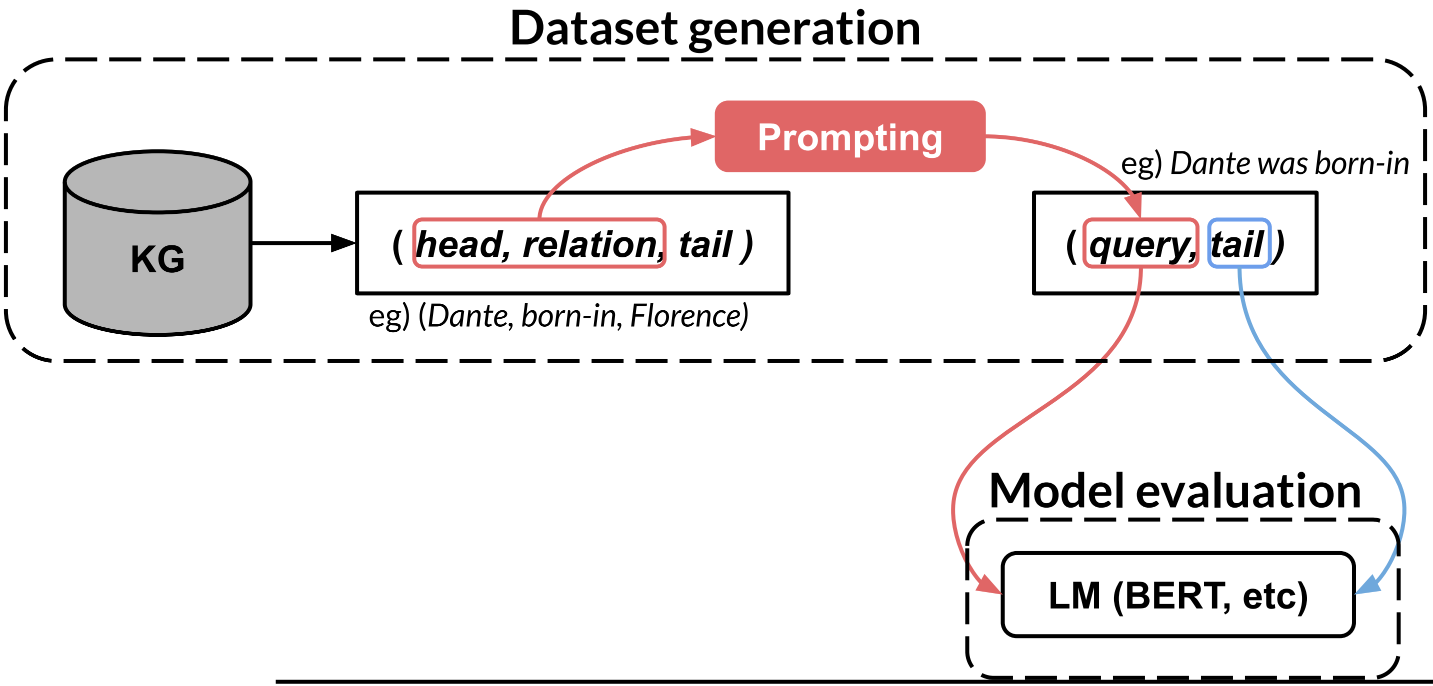
[Petroni, et al. "Language models as knowledge bases?"]

Dataset

- Dataset requirements:
  - Each entry has (*prompt*, *answer*)
    - eg) (*Dante was born in*, *Florence*)
  - Answer should be single token
  - Query should represent relational knowledge given a head and a relation in a KG
    - eg) *Dante was born in* = (*Dante*, *born-in*)

Petroni, et al. "Language models as knowledge bases?"

Whole pipeline



[Petroni, et al. "Language models as knowledge bases?"]

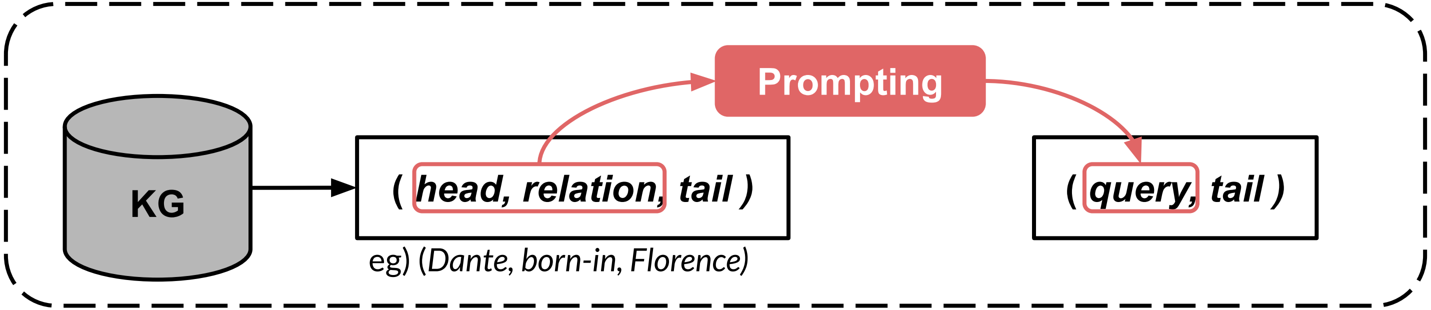
Results

Corpus	Relation	Statistics #Facts	#Rel	Baselines Freq	DrQA	KB RE_n	RE_o	LM Fs	Txl	Eb	E5B	Bb	BERT Large Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Petroni, et al. "Language models as knowledge bases?"

Effect of prompt type

Dataset generation



eg) (*Dante, born-in, Florence*)

What's the best prompt? 🤔

- *Dante was born-in Florence.*
- *Florence is where Dante was born*
- *Dante was born-in Florence, Italy.*
- *etc*

[Jiang, et al. "How can we know what language models know?"](#)

Prompt ensembling/selection

- Prompting methods
  - Manual
  - Mined: Frequency in a large corpus
  - Paraphrased: Back-translation
- Ensembling prompt
  - Optimization over training set
- Data:
  - Test: T-Rex
  - Training: Wikidata

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	<u>Microsoft</u>	-1.77	<u>Microsoft</u>	-2.23
2	<u>Microsoft</u>	-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

[Jiang, et al. "How can we know what language models know?"]

Improve manual prompt

Averaging Ensembling

Prompts	Top1	Top3	Top5	Opti.	Oracle
*BERT-base (**Man**=31.1)*					
Mine	30.7	32.7	31.2	36.9	45.1
Mine+Man	31.9	34.5	33.8	**38.1**	47.9
Mine+Para	30.7	33.0	33.7	33.6	45.0
Man+Para	*34.1*	35.8	36.6	37.3	47.9
*BERT-large (**Man**=32.3)*					
Mine	34.4	33.8	33.1	40.4	47.9
Mine+Man	*36.0*	38.6	37.1	**41.9**	50.8
Mine+Para	32.1	35.0	36.1	37.0	47.3
Man+Para	35.9	37.3	38.0	38.8	50.0

Jiang, et al. "How can we know what language models know?"

Mined prompts

ID	Relations	Manual Prompts	Mined Prompts	Acc. Gain
P140	religion	x is affiliated with the y religion	x who converted to y	+60.0
P159	headquarters location	The headquarter of x is in y	x is based in y	+4.9
P20	place of death	x died in y	x died at his home in y	+4.6
P264	record label	x is represented by music label y	x recorded for y	+17.2
P279	subclass of	x is a subclass of y	x is a type of y	+22.7
P39	position held	x has the position of y	x is elected y	+7.9

[Jiang, et al. "How can we know what language models know?"]

Paraphrased prompts

ID	Modifications	Acc. Gain
P413	x plays in→at y position	+23.2
P495	x was created→made in y	+10.8
P495	x was→is created in y	+10.0
P361	x is a part of y	+2.7
P413	x plays ~~in~~ y position	+2.2

Jiang, et al. "How can we know what language models know?"

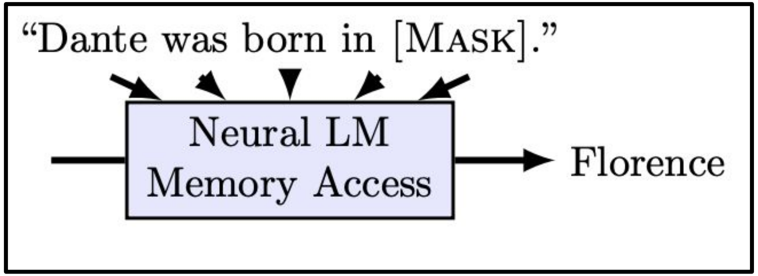
Ensembling weight

ID	Relations	Prompts and Weights	Acc. Gain
P127	owned by	x is owned by y .485 x was acquired by y .151 x division of y .151	+7.0
P140	religion	x who converted to y .615 y tirthankara x .190 y dedicated to x .110	+12.2
P176	manufacturer	y introduced the x .594 y announced the x .286 x attributed to the y .111	+7.0

Test	BERT-base		BERT-large	
Train	base	large	large	base
mine	36.9	36.6	40.4	39.0
mine+man	38.1	38.5	41.9	40.0
mine+para	33.6	33.7	37.0	35.3
man+para	37.3	35.6	38.8	37.5

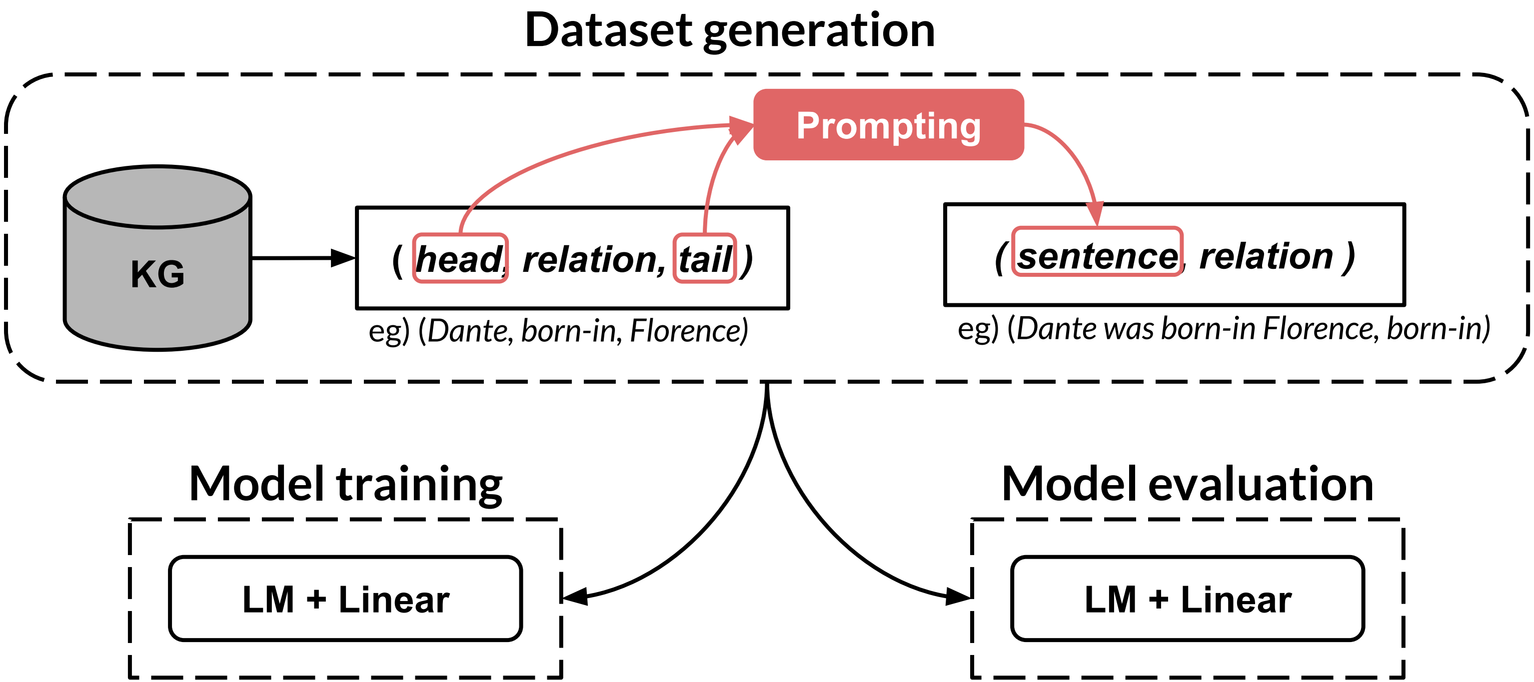
[Jiang, et al. "How can we know what language models know?"](#)

Issue with link prediction



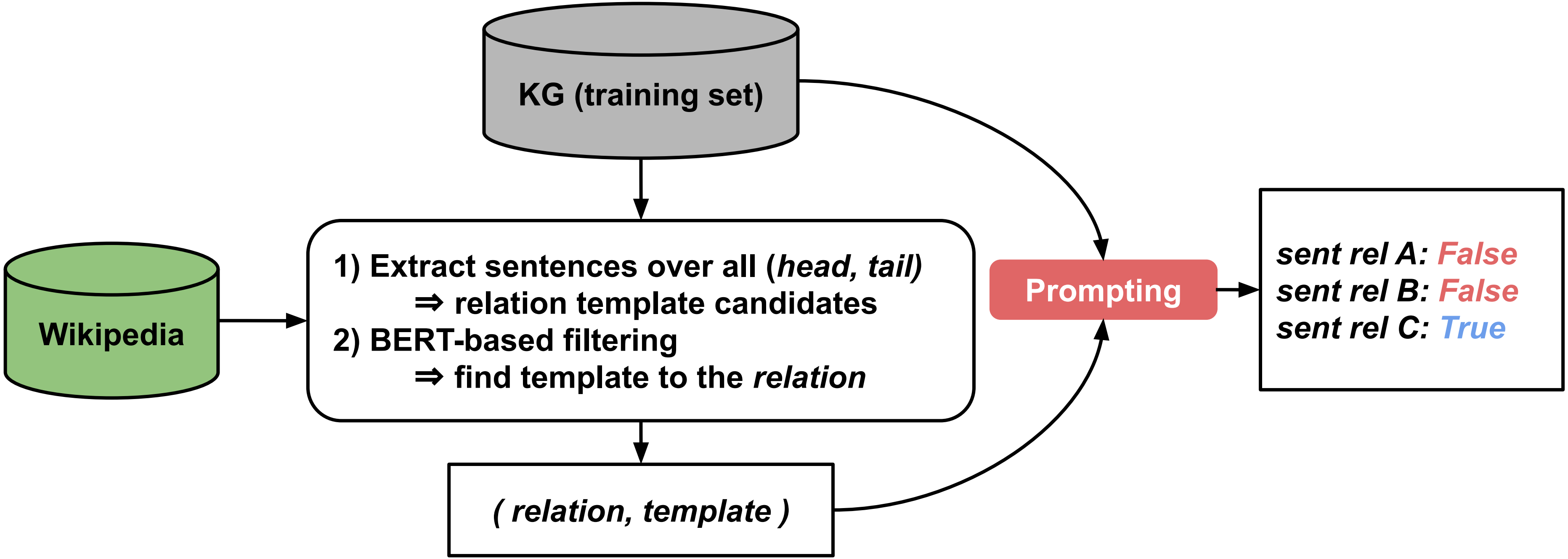
- Spurious correlation among subject and object
  - *Birds cannot [MASK] → fly*, Kassner and Schütze, 2020
  - *The capital of Macintosh is [MASK] → apple*, *Bouraoui, et al. 2019*
- Heuristics on surface form
  - BERT is very good at IR, Petroni, et al. 2020

Relation classification



[Bouraoui, et al. "Inducing relational knowledge from BERT."](#)

Template search and prompt



[Bouraoui, et al. "Inducing relational knowledge from BERT."]()

Classification result

word embedding

	Google			DiffVec			BATS		
	pr	rec	f1	pr	rec	f1	pr	rec	f1
SVM_{glove}	45.7	70.2	55.3	32.7	52.7	40.3	42.3	55.6	48.0
SVM_{sg}	49.4	68.9	57.5	38.5	47.2	42.4	42.9	61.3	50.4
Trans_{glove}	76.9	72.5	74.6	39.6	59.6	47.5	53.4	65.6	58.8
Trans_{sg}	73.1	**74.3**	73.6	47.3	**72.6**	**57.2**	**63.1**	**70.6**	**66.6**
$\mathrm{BERT}^{max}{}_{50}$	85.2	67.1	75.0	58.1	43.4	49.6	57.3	36.5	44.5
$\mathrm{BERT}^{max}{}_{100}$	**86.8**	69.3	**77.0**	59.5	46.7	52.8	60.3	41.7	49.5
$\mathrm{BERT}^{max}{}_{1000}$	75.8	58.2	65.8	52.9	40.3	45.7	56.3	37.1	44.7
$\mathrm{BERT}^{*}{}_{50}$	78.6	61.8	69.1	51.1	39.2	44.3	50.3	32.4	39.4
$\mathrm{BERT}^{*}{}_{100}$	79.4	63.7	70.6	**63.2**	47.8	54.4	59.2	44.5	50.8
$\mathrm{BERT}^{*}{}_{1000}$	76.9	51.0	61.3	53.1	38.5	44.6	57.6	35.3	43.7

language model

[Bouraoui, et al. "Inducing relational knowledge from BERT."]

Result breakdown: Google

	Google	Trans_{sg}	SVM_{sg}	BERT_{100}^{max}
Morphological	gram1-adj-to-adv	**63.5**	51.2	49.9
	gram2-opposite	59.2	49.6	**68.5**
	gram3-comparative	**79.7**	62.1	78.4
	gram4-superlative	**88.3**	49.4	86.6
	gram5-present-participle	**70.1**	56.1	68.9
	gram6-nationality-adj	63.8	58.3	**79.6**
	gram7-past-tense	**80.1**	54.2	67.6
	gram8-plural	**72.9**	68.9	48.8
	gram9-plural-verbs	**69.4**	51.1	65.8
Semantic	currency	82.3	60.1	**93.6**
	capital-common-countries	82.3	73.4	**91.2**
	capital-world	78.1	62.0	**89.5**
	family	72.7	52.3	**88.2**
	city-in-state	68.4	57.2	**79.6**

[Bouraoui, et al. "Inducing relational knowledge from BERT."]()

Result breakdown: DiffVec

	DiffVec	Trans_{sg}	SVM_{sg}	BERT^{max}_{100}
Attr.	Action:ObjectAttribute	19.2	20.1	**35.2**
	Object:State	56.2	32.1	**58.0**
	Object:TypicalAction	25.3	35.4	**49.0**
Causality	Action/Activity:Goal	31.9	29.3	**57.1**
	Agent:Goal	43.5	36.7	**53.9**
	Cause:CompensatoryAction	59.1	46.8	**63.4**
	Cause:Effect	63.4	42.4	**64.0**
	EnablingAgent:Object	34.3	45.5	**58.7**
	Instrument:Goal	56.8	41.2	**60.5**
	Instrument:IntendedAction	62.9	39.2	**68.8**
	Prevention	70.1	53.2	**72.1**
Lexical	Collective noun	**55.6**	40.8	38.1
	Hyper	**73.6**	41.5	54.3
	Lvc	75.0	**75.6**	37.4
	Mero	**64.6**	41.4	47.5
Commonsense	Event	50.2	39.8	**57.8**
	Concealment	42.1	32.4	**52.7**
	Expression	**80.3**	52.3	79.3
	Knowledge	70.1	51.4	**72.4**
	Plan	56.5	32.3	**62.3**
	Representation	48.2	39.7	**50.1**
	Sign:Significant	38.1	30.2	**41.1**
	Attachment	36.4	41.0	**52.9**
	Contiguity	61.2	32.8	**70.8**
	Item:Location	28.1	32.1	**54.2**
	Loc:Action/Activity	74.8	51.3	**77.4**
	Loc:Instr/AssociatedItem	42.0	44.9	**69.0**
	Loc:Process/Product	47.2	56.6	**64.3**
	Sequence	62.8	50.2	**74.9**
	Time:Action/Activity	57.2	53.7	**59.1**
Morphological	Noun Singplur	**53.0**	38.5	33.5
	Prefix re	**71.5**	30.2	19.6
	Verb 3rd	**97.0**	38.4	20.3
	Verb 3rd Past	**95.3**	32.2	21.9
	Verb Past	**82.1**	61.3	26.6
	Vn-Deriv	**75.5**	63.1	25.0

Bouraoui, et al. "Inducing relational knowledge from BERT."

Result breakdown: BATS

	BATS	Trans_{sg}	SVM_{sg}	BERT^{max}_{100}
Morphological	Regular plurals	**76.3**	40.8	35.0
	Plurals - orth. changes	**76.0**	48.1	25.5
	Comparative degree	**76.2**	47.5	50.2
	Superlative degree	**82.1**	59.5	53.3
	Infinitive: 3Ps.Sg	**82.0**	59.8	25.5
	Infinitive: participle	**79.4**	62.7	33.3
	Infinitive: past	**70.9**	52.0	35.1
	Participle: 3Ps.Sg	**78.3**	62.9	29.9
	Participle: past	**76.3**	56.7	36.7
	3Ps.Sg: past	**86.4**	65.8	25.9
	Noun+less	**62.5**	43.8	26.6
	Un+adj	**71.2**	40.5	28.8
	Adj+ly	**73.0**	39.8	35.5
	Over+adh./Ved	**71.1**	41.5	36.7
	Adj+ness	**72.5**	53.6	30.5
	Re+verb	**75.1**	56.8	33.9
	Verb+able	**73.8**	55.3	25.4
	Verb+er	**60.2**	53.3	42.3
	Verb+ation	**58.9**	46.6	28.8
	Verb+ment	**60.6**	48.1	40.7

Lexical	Hypernyms animals	63.6	64.5	**71.8**
	Hypernyms misc	78.1	56.2	**78.8**
	Hyponyms misc	54.6	50.9	**61.3**
	Meronyms substance	**53.1**	37.8	50.4
	Meronyms member	**70.2**	57.1	56.6
	Meronyms part-whole	49.5	52.3	**58.2**
	Synonyms intensity	46.7	35.6	**50.8**
	Synonyms exact	41.3	29.9	**48.7**
	Antonyms gradable	**49.3**	51.9	48.5
	Antonyms binary	49.6	33.3	**54.5**
Encyclopedic	Capitals	68.6	52.1	**73.2**
	Country:language	62.8	53.5	**69.5**
	UK city: county	61.6	48.0	**71.8**
	Nationalities	83.3	61.5	**84.4**
	Occupation	61.8	49.9	**72.6**
	Animals young	51.2	50.7	**68.2**
	Animals sounds	60.1	45.9	**63.1**
	Animals shelter	45.8	45.2	**63.3**
	thing:color	75.6	58.9	**76.5**
	male:female	76.9	49.3	**79.0**

[Bouraoui, et al. "Inducing relational knowledge from BERT."](#)

Recap

Link prediction: finetuning-free, but other factors

- *Petroni, et al. "Language models as knowledge bases?" 2019*
- *Jiang, et al. "How can we know what language models know?" 2019*

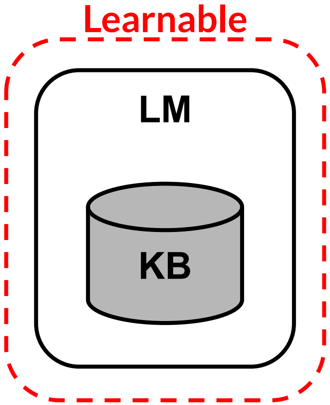
Relation classification: relation evaluation, but finetuning

- *Bouraoui, et al. "Inducing relational knowledge from BERT." 2019*

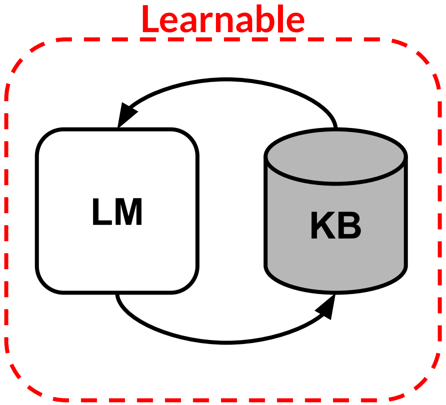
Limitation/Open issue

- Object with multiple tokens
- Better template
- More dataset
- Effect on other tasks
- etc...

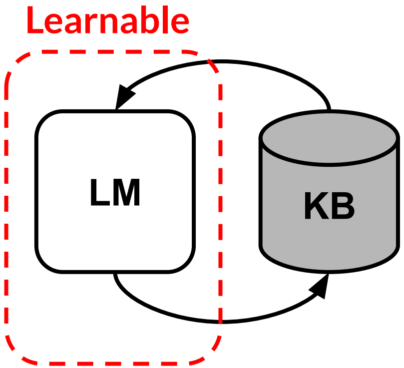
Thanks for listening 🎉

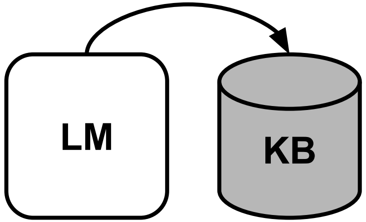


Related Topic



1. KB augmented LM
    - Latent Relation LMs, KnowBERT, COMET
2. LM training with KB
    - RAG, REALM
3. LM inference with KB
    - kNN-LM, BERT-kNN, IR+BERT
4. KB completion
    - Commonsense KB completion, LMs are open KG





Comment given to the talk

- We need to differentiate LM as a language generator and fact retriever
- BPE subword handling
- More complex reasoning