K-BERT: Enabling Language Representation with Knowledge Graph

Weijie Liu,[1] Peng Zhou,[2] Zhe Zhao,[2] Zhiruo Wang,[3] Qi Ju,[2,*]
Haotang Deng,[2] Ping Wang[1,*]
[1]Peking University, Beijing, China
[2]Tencent Research, Beijing, China
[3]Beijing Normal University, Beijing, China
{dataliu, pwang}@pku.edu.cn, SherronWang@gmail.com,
{rickzhou, nlpzhe, zhiruowang, damonju, haotangdeng}@tencent.com

Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV[*] Nelson F. Liu[†§] Matthew E. Peters[§]
Matt Gardner[§] Sameer Singh[*]
[*] University of California, Irvine, CA, USA
[†] University of Washington, Seattle, WA, USA
[§] Allen Institute for Artificial Intelligence, Seattle, WA, USA
{rlogan, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nfliu@cs.washington.edu

PRETRAINED ENCYCLOPEDIA: WEAKLY SUPERVISED KNOWLEDGE-PRETRAINED LANGUAGE MODEL

Wenhan Xiong[†], Jingfei Du[§], William Yang Wang[†], Veselin Stoyanov[§],
[†] University of California, Santa Barbara
[§] Facebook AI
{xwhan, william}@cs.ucsb.edu, {jingfeidu, ves}@fb.com

Language Models as Knowledge Bases?

Fabio Petroni[1] Tim Rocktäschel[1,2] Patrick Lewis[1,2] Anton Bakhtin[1]
Yuxiang Wu[1,2] Alexander H. Miller[1] Sebastian Riedel[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com

Nilesh & Tergel
2022.05.27

K-BERT

Background

- Pre-trained language models capture general language representation as opposed to the human experts

- BERT, GPT, and XLNet were pre-trained over open-domain corpora

- Knowledge Graphs (KG) will equip the model with domain knowledge, enhancing the model's performance over domain-specific tasks

Challenges:

- Heterogeneous Embedding Space (HES): In general, the embedding vectors of words in text and entities in KG are obtained in separate ways, making their vector-space inconsistent

- Knowledge Noise(KN): Too much knowledge incorporation may divert the sentence from its correct meanin g.

Solution:

- Knowledge-enabled Bidirectional Encoder Representation from Transformers (K-BERT) is capable of loading any pre-trained BERT models due to they are identical in parameters

- K-BERT can easily inject domain knowledge into the models by equipped with a KG without pre-training.

Methodology

Notation:

- sentence s = {w0, w1, w2, w3, ….., wn}   length = n
- English tokens vs Chinese tokens (wi)
- wi is an element in Vocab V of KG.
- KG is a collection of triples, ∈ = (wi, rj, wk)
- wi, wk are entities names
- rj is the relation between them

Model Architecture

- Four Modules:
  - Knowledge Layer
  - Embedding Layer
  - Seeing Layer
  - Mask-transformer

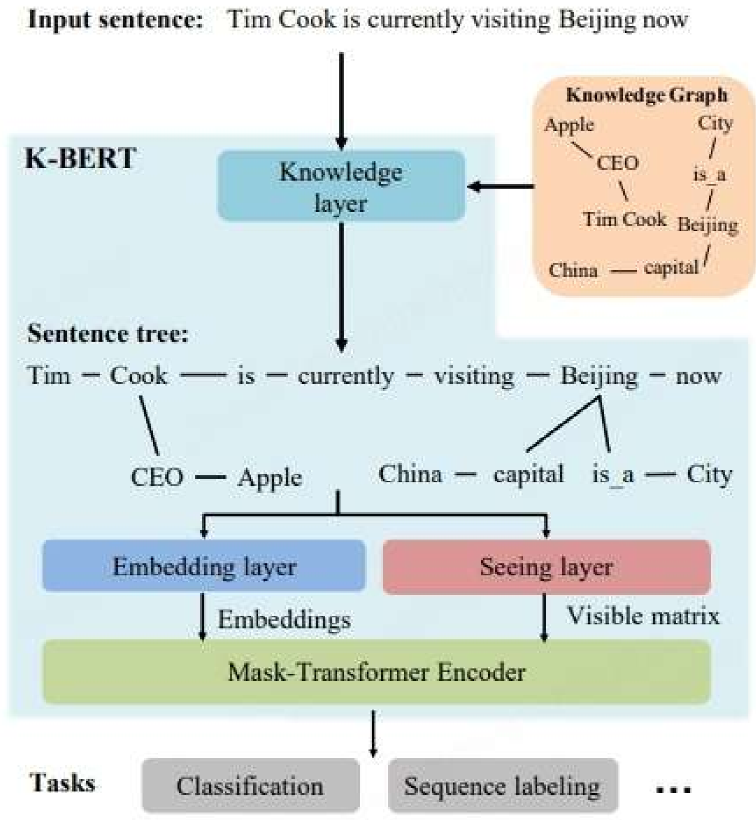


Figure 1: The model structure of K-BERT: Compared to other RL models, the K-BERT is equipped with an editable KG, which can be adapted to its application domain. For example, for electronic medical record analysis, we can use a medical KG to grant the K-BERT with medical knowledge.

Knowledge Layer

- **Input :** sentence s = {w0, w1, w2, …., wn} and KG K
- **Output :** sentence tree t = {w0, w1, …, wi{(ri0, wi0), …,(rik, wik)}, …, wn}

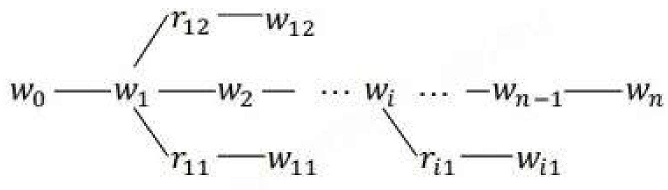


Figure 3: Structure of the sentence tree.

- **K-Query :** all the entity names involved (an entity that's identified) in the sentence s are selected out to quer y their corresponding triples from K.
- K-Query can be formulated as, E = K Query(s, K),
- where E = {(wi , ri0, wi0), …,(wi , rik, wik)} is a collection of the corresponding triples

- **K-Inject :** injects the queried E into the sentence s by stitching the triples in E to their corresponding positio n, and generates a sentence tree t

Methodology

KISTI SINCE 1962

Embedding Layer

- convert the sentence tree into an embedding representation that can be fed into the Mask-Transformer
- Similar to BERT, only the input is a sentence tree instead of a token sequence

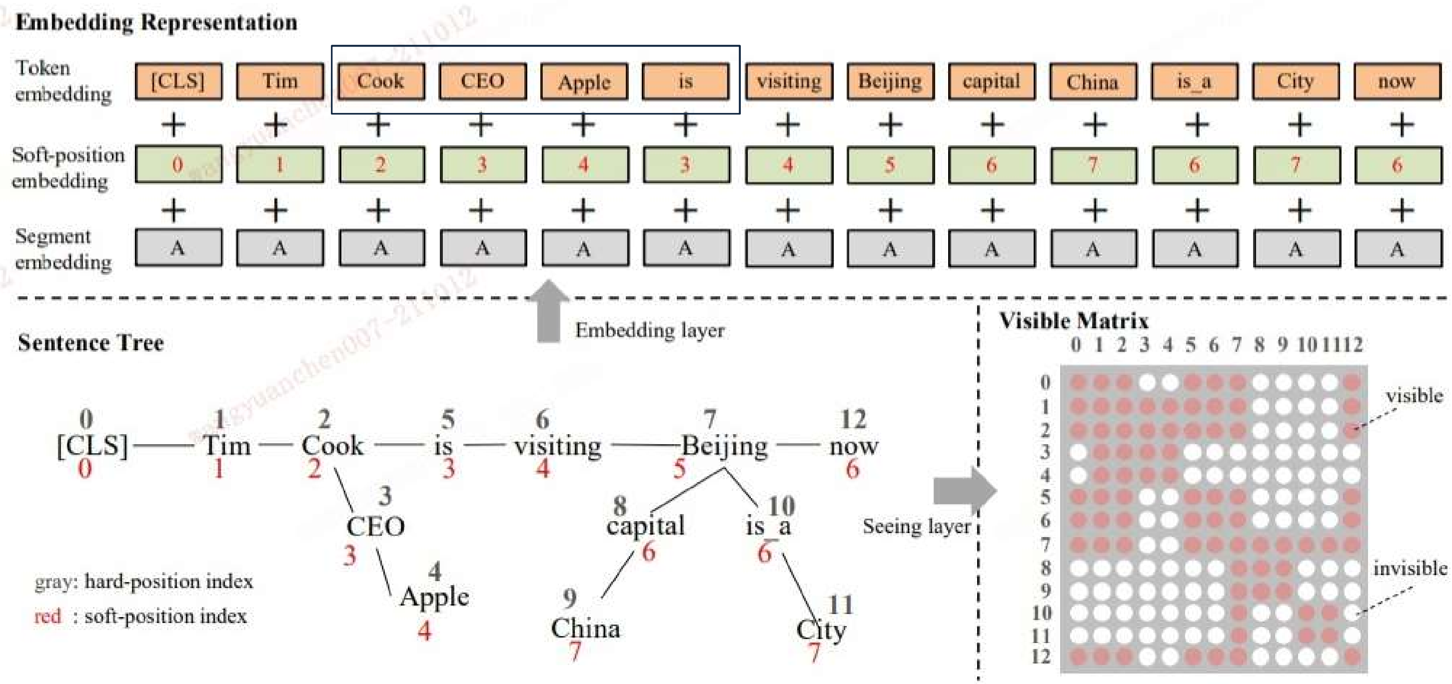
Token Embedding

- Tokens require rearrangement before embedding operation.
- Tokens in the branch are inserted after the corresponding node, while subsequent tokens are moved backw ard.

Soft-position Embedding

Segment embedding

Similar to BERT uses segmentation embedding to identify differently sentences when multiple sentences are included.



Seeing Layer

- biggest difference between K-BERT and BERT, and also what makes this method so effective
- To tackle KN (Knowledge Noise) issue, the authors propose a visible matrix M to limit the visible area of ea ch token so that the additional information extracted from KG would not be visible to all tokens
- The visibility mechanism can be presented as a function: (hard position means to exclude the soft position)

$$M_{ij} = \begin{cases} 0 & w_i \ominus w_j \\ -\infty & w_i \oslash w_j \end{cases} \tag{3}$$

where, $w_i \ominus w_j$ indicates that w_i and w_j are in the same branch, while $w_i \oslash w_j$ are not. i and j are the hard-position index.

Methodology

KISTI SINCE 1962

Mask Transformer

- Mask-Transformer can limit the self-attention region according to M
- As BERT, they denote the number of layers (i.e., mask-self-attention blocks) as L, the hidden size as H, and the number of mask-self-attention heads as A
- Formally, the mask-self-attention is:

$$Q^{i+1}, K^{i+1}, V^{i+1} = h^i W_q,\ h^i W_k,\ h^i W_v, \tag{4}$$

$$S^{i+1} = softmax\left(\frac{Q^{i+1} {K^{i+1}}^\top + M}{\sqrt{d_k}}\right), \tag{5}$$

$$h^{i+1} = S^{i+1} V^{i+1}, \tag{6}$$

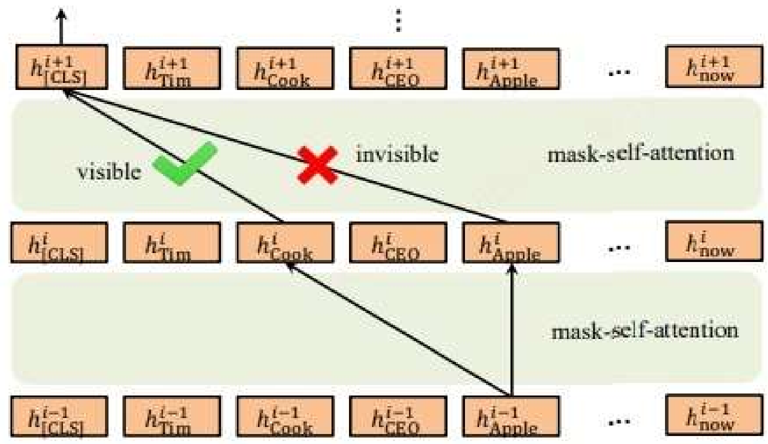


Figure 4: Illustration of the Mask-Transformer, which is a stack of multiple mask-self-attention blocks.

Results Discussions

Table 1: Results of various models on sentence classification tasks on open-domain tasks (*Acc.* %)

Models\Datasets	Book_review		Chnsenticorp		Shopping		Weibo		XNLI		LCQMC	
	Dev	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*
Pre-trainied on WikiZh by Google.												
Google BERT	88.3	**87.5**	93.3	94.3	96.7	96.3	98.2	98.3	76.0	75.4	88.4	86.2
K-BERT (HowNet)	88.6	87.2	**94.6**	**95.6**	**97.1**	**97.0**	98.3	98.3	**76.8**	**76.1**	**88.9**	86.9
K-BERT (CN-DBpedia)	**88.6**	87.3	93.9	95.3	96.6	96.5	98.3	98.3	76.5	76.0	88.6	**87.0**
Pre-trained on WikiZh and WebtextZh by us.												
Our BERT	**88.6**	87.9	94.8	95.7	96.9	**97.1**	98.2	98.2	77.0	76.3	89.0	86.7
K-BERT (HowNet)	88.5	87.4	**95.4**	95.6	96.9	96.9	98.3	**98.4**	**77.2**	**77.0**	**89.2**	**87.1**
K-BERT (CN-DBpedia)	88.8	87.9	95.0	**95.8**	**97.1**	97.0	98.3	98.3	76.2	75.9	89.0	86.9

Results Discussions

KISTI SINCE 1962

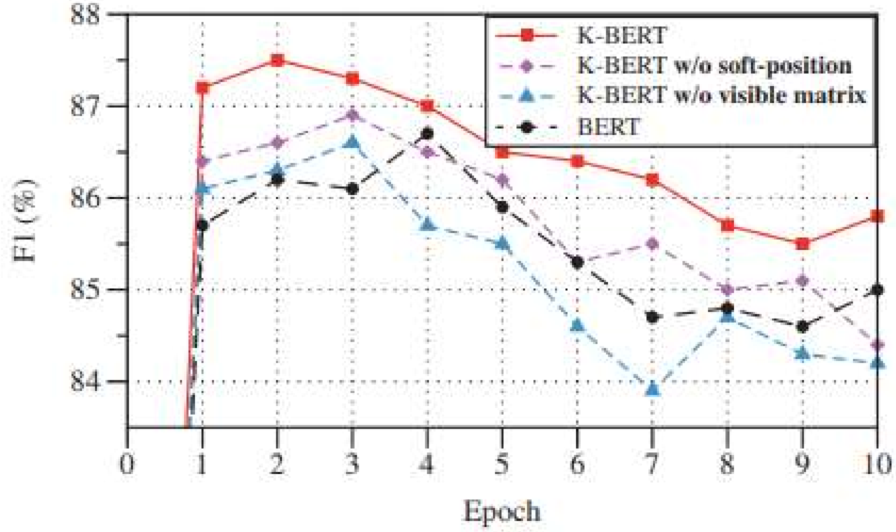
Table 2: Results of various models on NLPCC-DBQA (MRR %) and MSRA-NER ($F1$ %).

Models\Datasets	NLPCC-DBQA		MSRA-NER	
	Dev	*Test*	*Dev*	*Test*
Pre-trained on WikiZh by Google.				
Google BERT	93.4	93.3	94.5	93.6
K-BERT (HowNet)	93.2	93.1	95.8	94.5
K-BERT (CN-DBpedia)	**94.5**	**94.3**	**96.6**	**95.7**
Pre-trained on WikiZh and WebtextZh by us.				
Our BERT	93.3	93.6	95.7	94.6
K-BERT (HowNet)	93.2	93.1	96.3	95.6
K-BERT (CN-DBpedia)	**93.6**	**94.2**	**96.4**	**95.6**

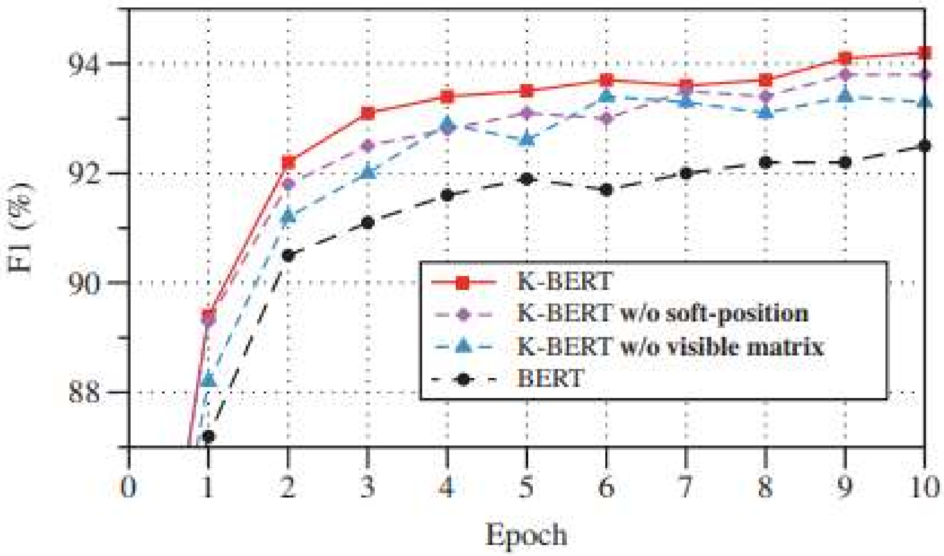
Table 3: Results of various models on specific-domain tasks (%).

Models\Datasets	Finance_Q&A			Law_Q&A			Finance_NER			Medicine_NER		
	P.	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*
Pre-trained on WikiZh by Google.												
Google BERT	81.9	86.0	83.9	83.1	90.1	86.4	84.8	87.4	86.1	91.9	93.1	92.5
K-BERT (HowNet)	83.3	84.4	83.9	83.7	91.2	87.3	86.3	89.0	**87.6**	93.2	93.3	93.3
K-BERT (CN-DBpedia)	81.5	88.6	**84.9**	82.1	93.8	**87.5**	86.1	88.7	87.4	93.9	93.8	93.8
K-BERT (MedicalKG)	-	-	-	-	-	-	-	-	-	94.0	94.4	**94.2**
Pre-trained on WikiZh and WebtextZh by us.												
Our BERT	82.1	86.5	84.2	83.2	91.7	87.2	84.9	87.4	86.1	91.8	93.5	92.7
K-BERT (HowNet)	82.8	85.8	84.3	83.0	92.4	87.5	86.3	88.5	87.3	93.5	93.8	93.7
K-BERT (CN-DBpedia)	81.9	87.1	**84.4**	83.1	92.6	**87.6**	86.3	88.6	**87.4**	93.9	94.3	94.1
K-BERT (MedicalKG)	-	-	-	-	-	-	-	-	-	94.1	94.3	**94.2**

Results Discussions



(a) Law_Q&A



(b) Medicine_NER

Conclusion

- After a presentation of model performance in different open-domain & specific domains (e.g. finance, law) tasks, the overall investigation reveals promising results in twelve NLP tasks.

- K-BERT significantly outperforms BERT, which demonstrates that K-BERT is an excellent choice for solving the knowledge-driven problems that require experts.

- It can be concluded that the soft-position and the visible matrix can make K-BERT more robust to KN interference and thus make more efficient use of knowledge.

USING KNOWLEDGE GRAPHS FOR FACT-AWARE LANGUAGE MODELING

Background

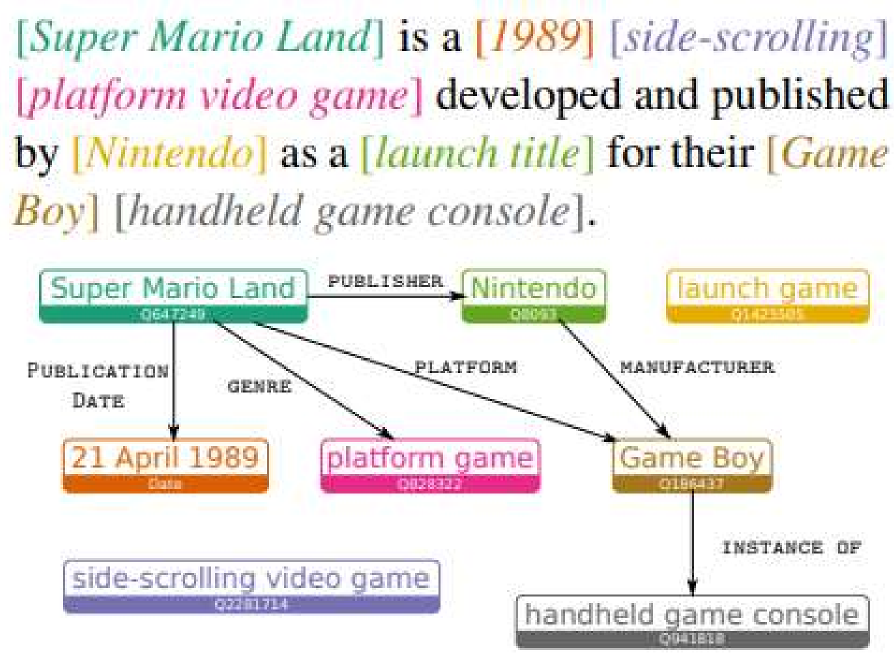
- LM should generate syntactically coherent as well as factually correct sentences

- The clearest limitation of existing language models is that they, at best, can only memorize facts observed during training

Proposed Solution:

- KGLM, a neural language model with mechanisms for selecting and copying information from an external k nowledge graph

- It maintains a dynamically growing local knowledge graph



Model

Language Model: LSTM

$$p(x_t|x_{<t}) = \mathrm{softmax}(\mathbf{W}_h\mathbf{h}_t + \mathbf{b}),$$
$$\mathbf{h}_t = \mathrm{RNN}(\mathbf{h}_{t-1}, \mathbf{x}_{t-1}).$$

Knowledge Graph: KG = {(p, r, e) | p ∈ E, r ∈ R, e ∈ E}

p – parent entity
r – relationship
e – other entity

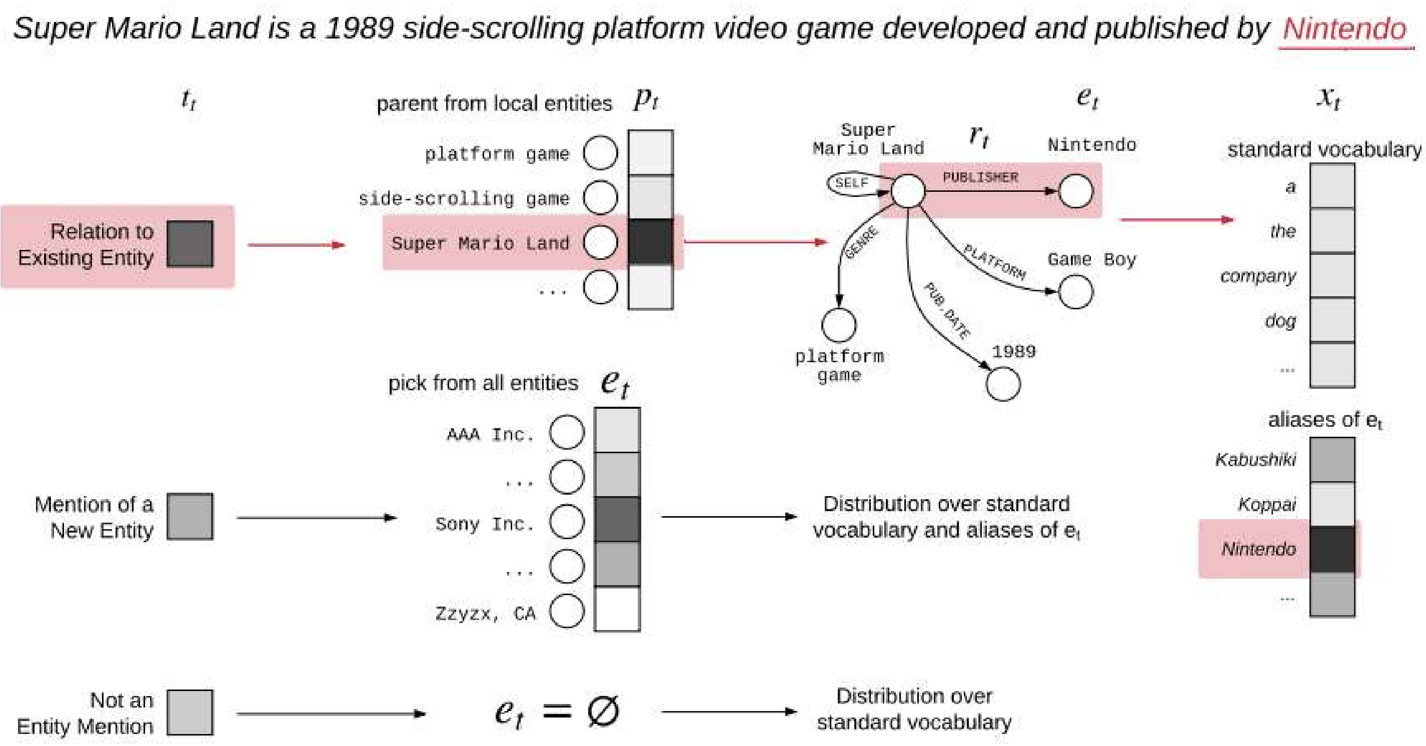
Caveats: integer value relations

Local Knowledge Graph: KG<t = {(p, r, e) | p ∈ E<t, r ∈ R, e ∈ E}

contains entities E<t and all facts they participate in

Generative KGLM

- KGLM will maintain a local knowledge graph containing all facts involving entities that have appeared in the context.
- It will grow the local knowledge graph with additional entities and facts to reflect the new entity
- We will compute, p(xt, Et |x<t, E<t)



Model

Marginalizing out the KG

- We will essentially marginalize the local knowledge graph to compute the probability of the tokens, i.e.

$$p(\mathbf{x}) = \sum_{\mathcal{E}} p(\mathbf{x}, \mathcal{E}).$$

Parameterizing the Distributions

- Now we compute the hidden state h_t $\quad h_t = [h_{t,x}; h_{t,P}; h_{tr}]$

- Token t_t is computed using a single-layer softmax over $h_{t,x}$ to predict one of {new, related, ∅}

- Picking an entity

$$p(e_t) = \mathrm{softmax}(\mathbf{v}_e \cdot (\mathbf{h}_{t,p} + \mathbf{h}_{t,r}))$$

$$p(p_t) = \mathrm{softmax}(\mathbf{v}_p \cdot \mathbf{h}_{t,p})$$

$$p(r_t) = \mathrm{softmax}(\mathbf{v}_r \cdot \mathbf{h}_{t,r})$$

- Rendering the entity

$$p(x_t = a_j) \propto \exp\left[\sigma\left(\left(\mathbf{h}'_{t,x}\right)^T \mathbf{W}_{\mathrm{copy}}\right) \mathbf{a}_j\right]$$

Model

- **Linked WikiText-2:** Solving the barrier of training data
- **Initial entity annotation:** human-provided links between Wikipedia article
- **Local knowledge graph:** iteratively creates a generative story for the entities using relations in the knowled ge graph as well as identifies new entities

Tokens	x_t	Super Mario Land	is	a	1989	side - scrolling	platform video game	developed
Mention type	t_t	new	$\emptyset$	$\emptyset$	related	new	related	$\emptyset$
Entity Mentioned	e_t	SML	$\emptyset$	$\emptyset$	04-21-1989	SIDE_SCROLL	PVG	$\emptyset$
Relation	r_t	$\emptyset$	$\emptyset$	$\emptyset$	pub date	$\emptyset$	genre	$\emptyset$
Parent Entity	p_t	$\emptyset$	$\emptyset$	$\emptyset$	SML	$\emptyset$	SML	$\emptyset$

x_t	and	published	by	Nintendo	as	a	launch title	for	their	Game Boy	handheld game console	.
t_t	$\emptyset$	$\emptyset$	$\emptyset$	related	$\emptyset$	$\emptyset$	new	$\emptyset$	$\emptyset$	related	related	$\emptyset$
e_t	$\emptyset$	$\emptyset$	$\emptyset$	NIN	$\emptyset$	$\emptyset$	LT	$\emptyset$	$\emptyset$	GAME_BOY	HGC	$\emptyset$
r_t	$\emptyset$	$\emptyset$	$\emptyset$	pub	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	R:manu / platform	instance of	$\emptyset$
p_t	$\emptyset$	$\emptyset$	$\emptyset$	SML	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	NIN / SML	GAME_BOY	$\emptyset$

Model

- **Dataset Statistics:**

	Train	Dev	Test
Documents	600	60	60
Tokens	2,019,195	207,982	236,062
Vocab. Size	33,558	-	-
Mention Tokens	207,803	21,226	24,441
Mention Spans	122,983	12,214	15,007
Unique Entities	41,058	5,415	5,625
Unique Relations	1,291	484	504

Table 2: ***Linked WikiText-2*** **Corpus Statistics.**

Results

Fact Completion

	AWD-LSTM	GPT-2	KGLM Oracle	KGLM NEL
nation-capital	0 / 0	**6 / 7**	0 / 0	0 / 4
birthloc	0 / 9	14 / 14	**94 / 95**	85 / 92
birthdate	0 / 25	8 / 9	**65 / 68**	61 / 67
spouse	0 / 0	**2** / 3	**2** / 2	1 / **19**
city-state	0 / 13	**62 / 62**	9 / 59	4 / 59
book-author	0 / 2	0 / 0	**61 / 62**	25 / 28
Average	0.0/8.2	15.3/15.8	**38.5/47.7**	29.3/44.8

Perplexity Results

	PPL	UPP
ENTITYNLM* (Ji et al., 2017)	85.4	189.2
EntityCopyNet*	76.1	144.0
AWD-LSTM (Merity et al., 2018)	74.8	165.8
KGLM*	**44.1**	**88.5**

Sentence Completion

	Input Sentence	Gold	GPT-2	KGLM
Both correct	Paris Hilton was born in ___	New York City	New	1981
	Arnold Schwarzenegger was born on ___	1947-07-30	July	30
KGLM correct	Bob Dylan was born in ___	Duluth	New	Duluth
	Barack Obama was born on ___	1961-08-04	January	August
	Ulysses is a book that was written by ___	James Joyce	a	James
GPTv2 correct	St. Louis is a city in the state of ___	Missouri	Missouri	Oldham
	Richard Nixon was born on ___	1913-01-09	January	20
	Kanye West is married to ___	Kim Kardashian	Kim	the
Both incorrect	The capital of India is ___	New Delhi	the	a
	Madonna is married to ___	Carlos Leon	a	Alex

Conclusion

- (KGLM), a neural language model that can access an external source of facts, encoded as a knowledge graph, in order to generate text.

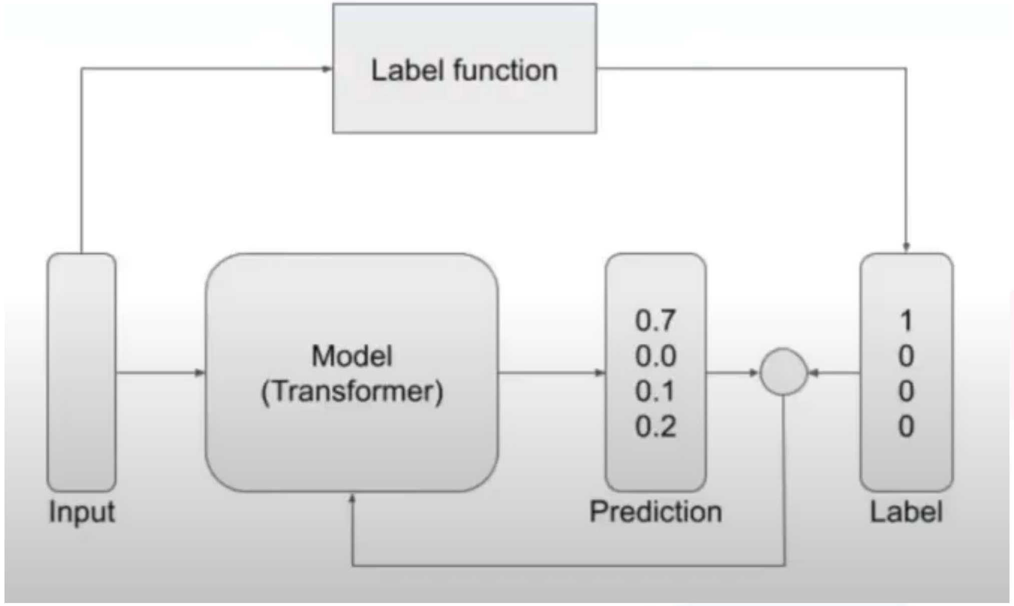
- KGLM is able to generate higher-quality, factually correct text that includes mentions of rare entities and specific tokens like numbers and dates.

PRETRAINED ENCYCLOPEDIA: WEAKLY SUPERVISED KNOWLEDGE-PRETRAINED LANGUAGE MODEL

Background – Weak Supervision

KISTI SINCE 1962

- Instead of a subject-matter expert (SME) hand-labelling high-quality data, all of which is very cost-prohibi tive, we can use other techniques that combine diverse sources of data, creating an approximation of lab els
- Labels are considered "weak" because they are noisy—i.e., the data measurements that the labels represe nt are not accurate and have a margin of error. The labels are also considered "weak" if they have additio nal information that does not directly indicate what we want to predict.



Key Contributions

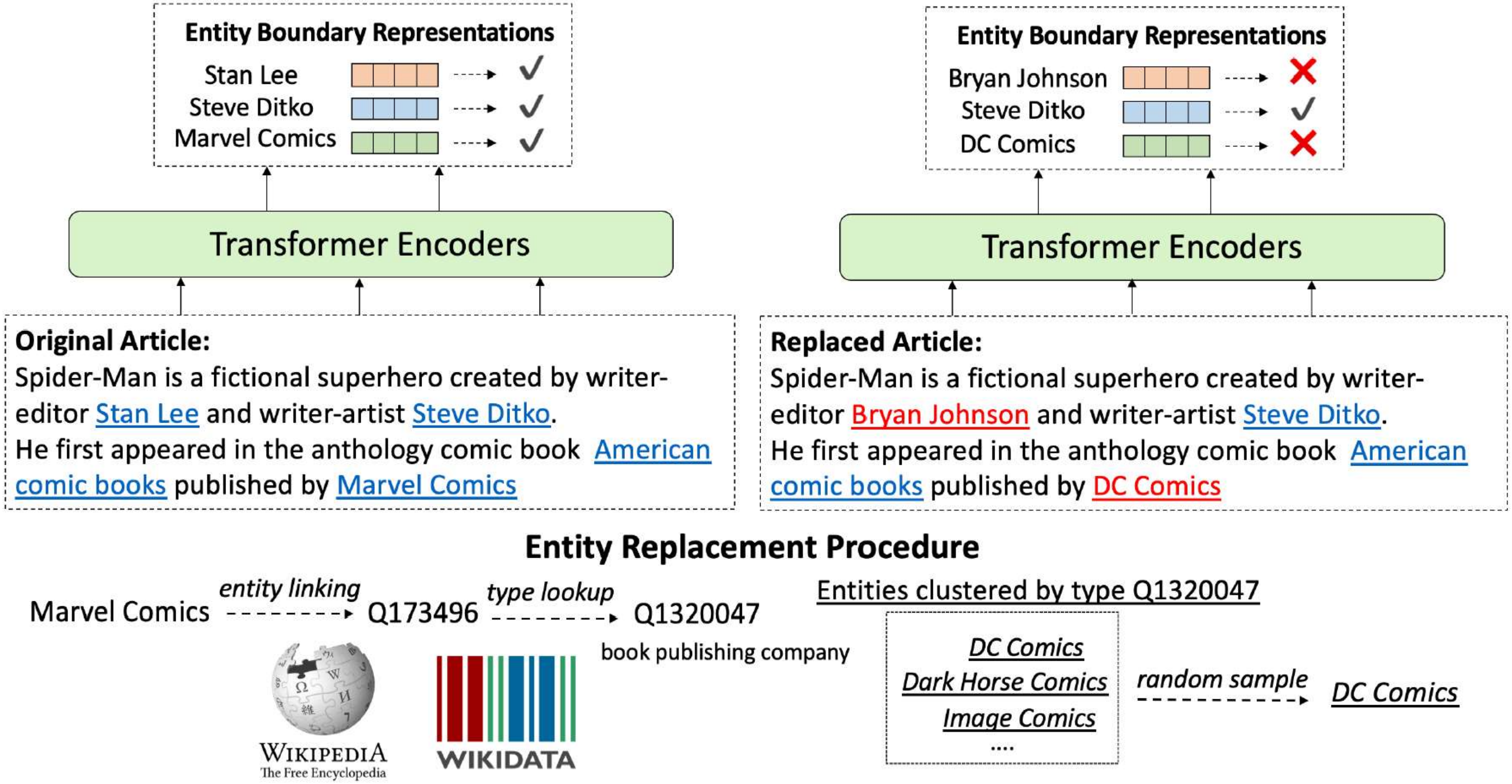
- Problem statement:
  - Existing pretraining objectives are usually defined at the token level and do not explicitly model entity-centric knowledge

- Objective
  - To test previous pretrained models' ability on encoding knowledge of common real-world entities
  - To improve the performance on knowledge about real-world entities from natural language text by proposing a new weakly supervised pretraining method

Original Article:
Spider-Man is a fictional superhero created by writer-editor Stan Lee and writer-artist Steve Ditko.
He first appeared in the anthology comic book American comic books published by Marvel Comics

Replaced Article:
Spider-Man is a fictional superhero created by writer-editor Bryan Johnson and writer-artist Steve Ditko.
He first appeared in the anthology comic book American comic books published by DC Comics

Methodology



Type-Constrained Entity Replacements for Knowledge Learning

Methodology

- Model architecture:
  - The same architecture as BERT base (12 Transformer layers)
  - They reimplemented and pretrained their own BERT
  - Concatenate the boundary words' representations + a linear layer + binary cross entropy
- Training objective:
  - Entity replacement objective
  - Masked language model objective (5% instead of 15%)
  - Restrict the masks to be outside the entity spans

Experiments – Zero-shot fact completion

- Dataset:
  - {*Paris, CapitalOf, France*} -> the capital of France is Paris
  - the capital of France is *[MASK]*

Table 1: Zero-Shot Fact Completion Results.

Relation Name	# of Candidates	# of Answers	Model **BERT-base**	**BERT-large**	**GPT-2**	**Ours**
HASCHILD (P40)	906	3.8	9.00	6.00	20.5	**63.5**
NOTABLEWORK (P800)	901	5.2	1.88	2.56	2.39	**4.10**
CAPITALOF (P36)	820	2.2	1.87	1.55	15.8	**49.1**
FOUNDEDBY (P112)	798	3.7	2.44	1.93	8.65	**24.2**
CREATOR (P170)	536	3.6	4.57	4.57	7.27	**9.84**
PLACEOFBIRTH (P19)	497	1.8	19.2	**30.9**	8.95	23.2
LOCATEDIN (P131))	382	1.9	13.2	52.5	21.0	**61.1**
EDUCATEDAT (P69)	374	4.1	9.10	7.93	11.0	**16.9**
PLACEOFDEATH (P20)	313	1.7	**43.0**	42.6	8.83	26.5
OCCUPATION (P106)	190	1.4	8.58	**10.7**	9.17	**10.7**
Average Hits@10	-	-	11.3	16.1	16.3	**28.9**

Experiments – Downstream tasks

- Dataset:

Table 2: Properties of the QA Datasets.

Dataset	Train	Valid	Test	Example Questions
WebQuestions	3778	-	2032	*Who plays Stewie Griffin on Family Guy?*
TriviaQA	87291	11274	10790	*What is the Japanese share index called?*
SearchQA	99811	13893	27247	*Hero several books 11 discover's wizard?*
Quasar-T	37012	3000	3000	*Which vegetable is a Welsh emblem?*

Table 4: Open-domain QA Results.

Model	WebQuestions		TriviaQA		Quasar-T		SearchQA	
	EM	F1	EM	F1	EM	F1	EM	F1
DrQA (Chen et al., 2017)	20.7	-	-	-	-	-	-	-
R^3 (Wang et al., 2018a)	-	-	50.6	57.3	42.3	49.6	57.0	63.2
DSQA (Lin et al., 2018)	18.5	25.6	48.7	56.3	42.2	49.3	49.0	55.3
Evidence Agg. (Wang et al., 2018b)	-	-	50.6	57.3	42.3	49.6	57.0	63.2
BERTserini (Yang et al., 2019a)	-	-	51.0	56.3	-	-	-	-
BERTserini+DS (Yang et al., 2019b)	-	-	54.4	60.2	-	-	-	-
ORQA (Lee et al., 2019)	**36.4**	-	45.0	-	-	-	-	-
Our BERT	29.2	35.5	48.7	53.2	40.4	46.1	57.1	61.9
Our BERT + Ranking score	32.2	38.9	52.1	56.5	43.2	49.2	60.6	65.9
WKLM	30.8	37.9	52.2	56.7	43.7	49.9	58.7	63.3
WKLM + Ranking score	34.6	41.8	**58.1**	**63.1**	**45.8**	**52.2**	**61.7**	**66.7**

Experiments – Entity typing

- Dataset:
  - FIGER

Table 5: Fine-grained Entity Typing Results on the FIGER dataset.

Model	Acc	Ma-F1	Mi-F1
LSTM + Hand-crafted (Inui et al., 2017)	57.02	76.98	73.94
Attentive + Hand-crafted (Inui et al., 2017)	59.68	78.97	75.36
BERT baseline (Zhang et al., 2019)	52.04	75.16	71.63
ERNIE (Zhang et al., 2019)	57.19	75.61	73.39
Our BERT	54.53	79.57	74.74
WKLM	**60.21**	**81.99**	**77.00**

Experiments – Ablation study

- THE EFFECT OF MASKED LANGUAGE MODEL LOSS

Table 6: Ablation Studies on Masked Language Model and Masking Ratios.

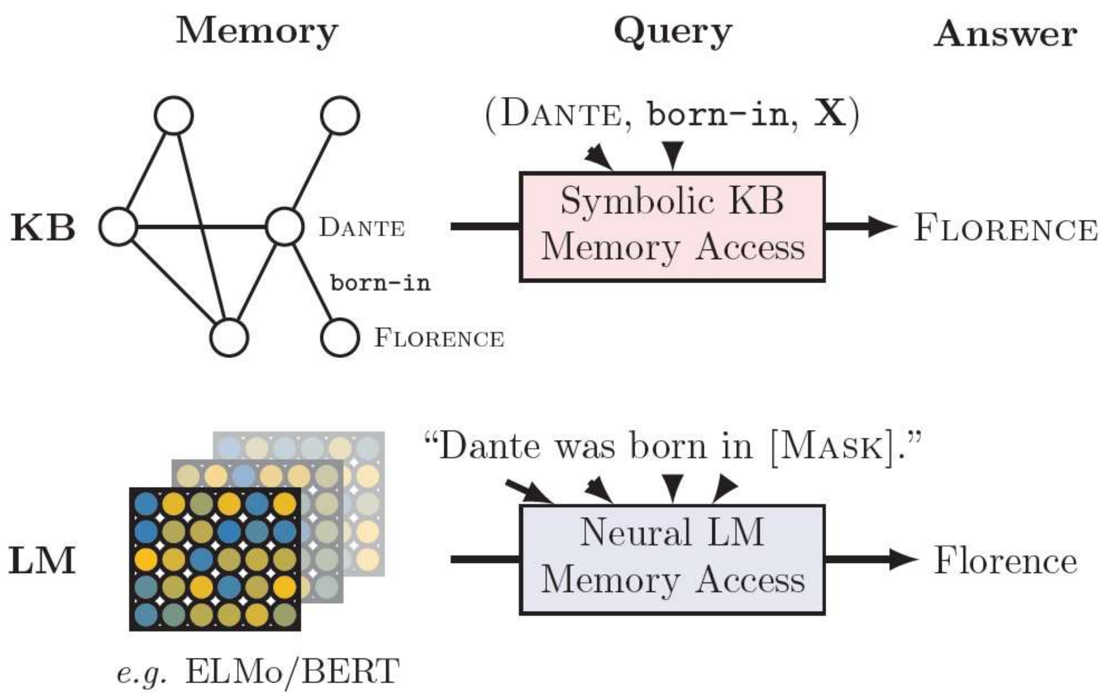
Model	SQuAD		TriviaQA		Quasar-T		FIGER
	EM	F1	EM	F1	EM	F1	Acc
Our BERT	83.4	90.5	48.7	53.2	40.4	46.1	54.53
WKLM	84.3	**91.3**	**52.2**	**56.7**	**43.7**	**49.9**	**60.21**
WKLM without MLM	80.5	87.6	48.2	52.5	42.2	48.1	58.44
WKLM with 15% masking	84.1	91.0	51.0	55.3	42.9	49.0	59.68
Our BERT + 1M MLM updates	**84.4**	91.1	52.0	56.3	42.3	48.2	54.17

Conclusion

- They proposed weakly supervised method to encourage pretrained language models to learn entity level knowledge
- It uses minimal entity information during pretraining and does not introduce a dditional computation, memory or architectural overhead for downstream task fine-tuning.
- The trained model demonstrates strong performance on a probing fact comple tion task and two entity-related NLP tasks

LANGUAGE MODELS AS KNOWLEDGE BASES?

Introduction



Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Key Contributions

- Problem statement:
  - How much relational knowledge do they (etc. BERT) store?
  - How does this different types of knowledge such as facts about entities, common sense, and general question answering?
  - How does their performance without fine-tuning compare to symbolic knowledge bases automatically extracted from text?

- Objective
  - To answer these questions by introducing the LAMA (LAnguage Model Analysis)

Methodology – The LAMA Probe

- Knowledge Source:

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	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

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naïve entity linking (RE_n), oracle entity linking (RE_o), fairseq-fconv (Fs), Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl) across the set of evaluation corpora.

Methodology – The LAMA Probe

- Models:

Model	Base Model	#Parameters	Training Corpus	Corpus Size
fairseq-fconv (Dauphin et al., 2017)	ConvNet	324M	WikiText-103	103M Words
Transformer-XL (large) (Dai et al., 2019)	Transformer	257M	WikiText-103	103M Words
ELMo (original) (Peters et al., 2018a)	BiLSTM	93.6M	Google Billion Word	800M Words
ELMo 5.5B (Peters et al., 2018a)	BiLSTM	93.6M	Wikipedia (en) & WMT 2008-2012	5.5B Words
BERT (base) (Devlin et al., 2018a)	Transformer	110M	Wikipedia (en) & BookCorpus	3.3B Words
BERT (large) (Devlin et al., 2018a)	Transformer	340M	Wikipedia (en) & BookCorpus	3.3B Words

Table 1: Language models considered in this study.

Methodology – The LAMA Considerations

- Manually Defined Templates
- Single Token
- Object Slots
- Intersections of Vocabularies

Results

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	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

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naïve entity linking (RE_n), oracle entity linking (RE_o), fairseq-fconv (Fs), Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl) across the set of evaluation corpora.

Results

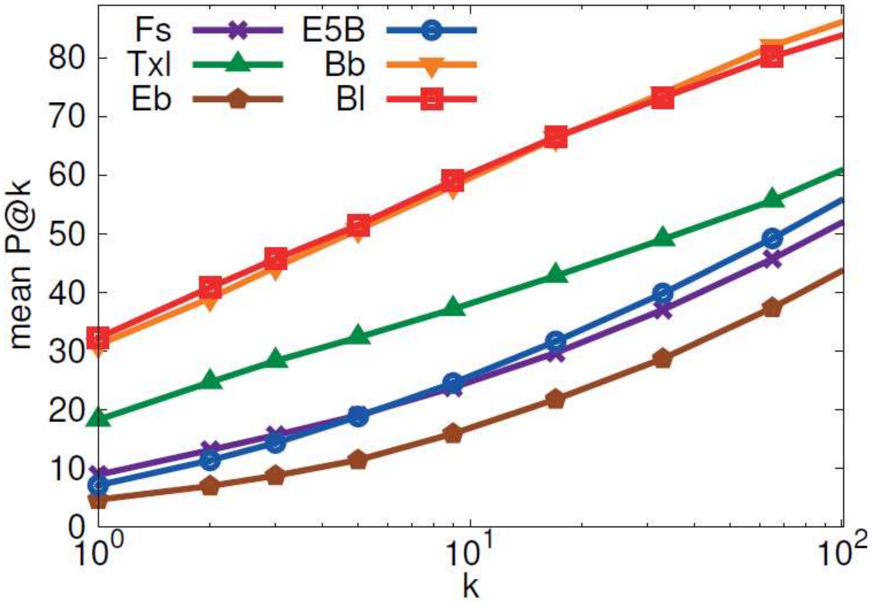


Figure 2: Mean P@k curve for T-REx varying k. Base-10 log scale for X axis.

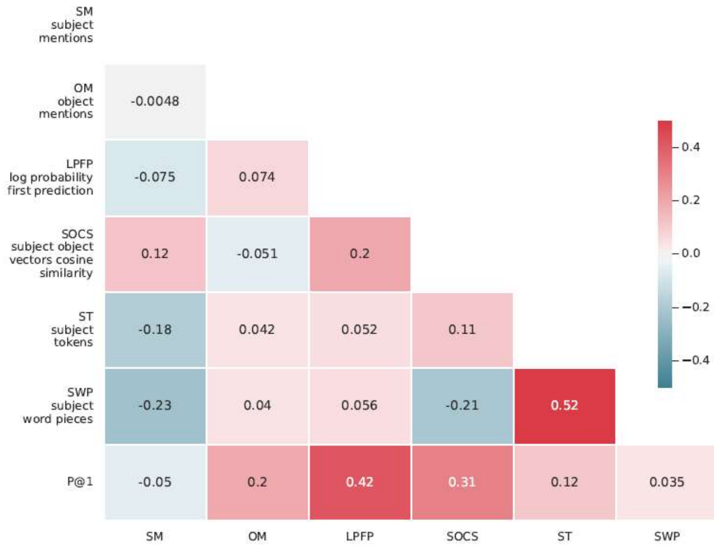


Figure 3: Pearson correlation coefficient for the P@1 of the BERT-large model on T-REx and a set of metrics: SM and OM refer to the number of times a subject and an object are mentioned in the BERT training corpus[4]respectively; LPFP is the log probability score associated with the first prediction; SOCS is the cosine similarity between subject and object vectors (we use spaCy[5]); ST and SWP are the number of tokens in the subject with a standard tokenization and the BERT WordPiece tokenization respectively.

Results

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence [-1.8]** , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]** , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , **dog [-2.4]** , cattle [-4.3] , sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]** , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , **midfielder [-2.4]** , forward [-2.4] , midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , **Hamburg [-7.5]** , Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]** , Breton [-3.3] , English [-3.8] , Dutch [-4.2] , German [-4.9]
	P54	Dani Alves plays with ____.	Barcelona	Santos [-2.4] , Porto [-2.5] , Sporting [-3.1] , Brazil [-3.3] , Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1] , journalist [-2.4] , teacher [-2.7] , doctor [-3.0] , physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2] , sulfur [-1.7] , **sodium [-2.5]** , zinc [-2.8] , salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3] , Conservative [-1.6] , Green [-2.4] , Liberal [-2.9] , **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0] , **Uganda [-3.2]** , Tanzania [-3.5] , China [-3.6] , Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]** , Nokia [-1.7] , Sony [-2.0] , Samsung [-2.6] , Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]** , Bermuda [-2.2] , Newfoundland [-2.5] , Alaska [-2.7] , Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0] , Intel [-2.3] , Microsoft [-2.5] , HP [-3.4] , Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6] , Latin [-1.9] , French [-2.4] , English [-3.0] , Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]** , Alberta [-2.8] , Yukon [-2.9] , Labrador [-3.4] , Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4] , Pope [-2.5] , **pope [-2.6]** , President [-3.1] , Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6] , BMG [-2.6] , Universal [-2.8] , **Capitol [-3.2]** , Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]** , Greenwich [-3.2] , Chelsea [-4.0] , Camden [-4.6] , Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1] , **ITV [-3.3]** , Global [-3.4] , Frontier [-4.1] , Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]** , Tamil [-2.1] , Telugu [-4.8] , English [-5.2] , Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2] , **Philippines [-3.6]** , February [-3.7] , December [-3.8] , Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1] , canal [-3.2] , toilet [-3.3] , stream [-3.6] , **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]** , fight [-1.8] , kill [-2.2] , die [-3.2] , hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7] , die [-1.7] , **laugh [-2.0]** , vomit [-2.6] , scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2] , cancer [-2.0] , **infection [-2.6]** , plague [-3.3] , fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8] , nests [-3.1] , **feathers [-3.2]** , died [-3.7] , eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5] , precision [-3.6] , registration [-3.8] , accuracy [-4.0] , **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7] , passing [-1.8] , precious [-2.9] , irrelevant [-3.2] , gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4] , human [-3.3] , **alive [-3.3]** , young [-3.6] , free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5] , useful [-2.5] , learned [-2.8] , combined [-3.9] , varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3] , fishing [-1.4] , bathing [-2.0] , **fish [-2.8]** , recreation [-3.1]

Table 3: Examples of generation for BERT-large. The last column reports the top five tokens generated together with the associated log probability (in square brackets).

Results

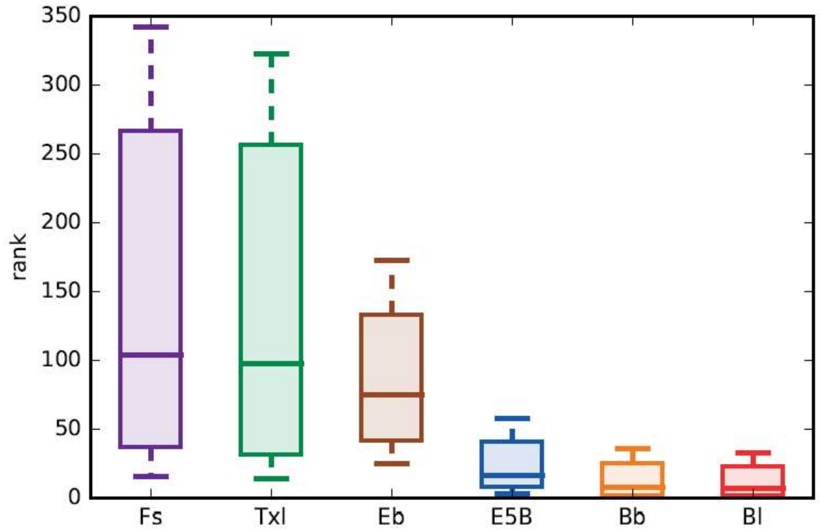


Figure 4: Average rank distribution for 10 different mentions of 100 random facts per relation in T-REx. ELMo 5.5B and both variants of BERT are least sensitive to the framing of the query but also are the most likely to have seen the query sentence during training.

Conclusion

KISTI SINCE 1962

- They presented a systematic analysis of the factual and common-sense knowledge in publicly available pretrained language models
- BERT-large is able to recall such knowledge better than its competitors