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[*]

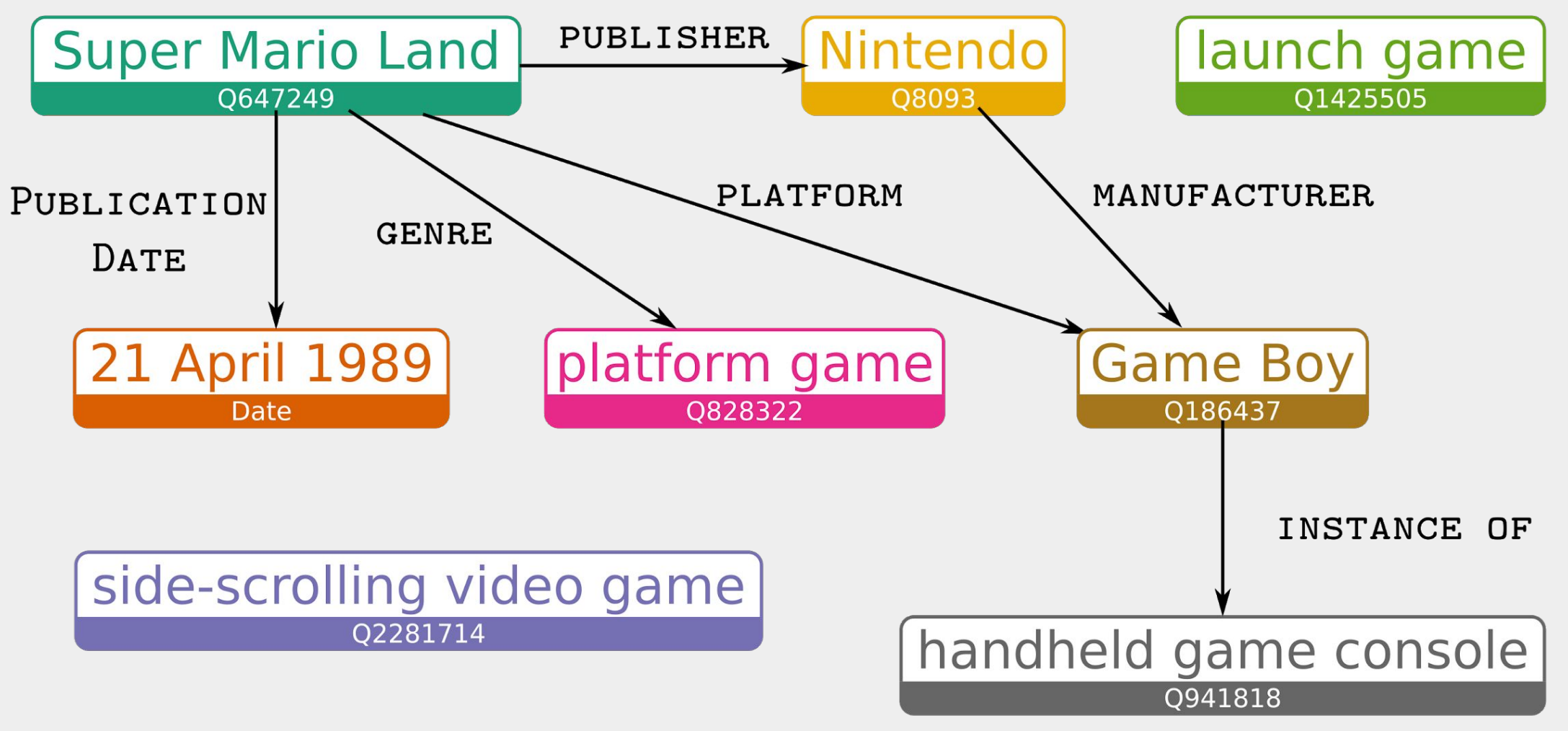
[*]University of California, Irvine, CA, USA, [†]University of Washington, Seattle, WA, USA, [§]Allen Institute for Artificial Intelligence, Seattle, WA, USA

Summary

- Traditional language models have limited ability to generate factually correct text.
- We introduce the *knowledge graph language model* (KGLM), a neural language model with mechanisms for generating information from a knowledge graph.
- We collect the *Linked WikiText-2* dataset, which aligns WikiText-2 to the Wikidata knowledge graph.
- Experiments show that the KGLM has better perplexity than AWD-LSTM-LM, and better fact-completion capabilities than GPT-2 small despite being trained on less data.

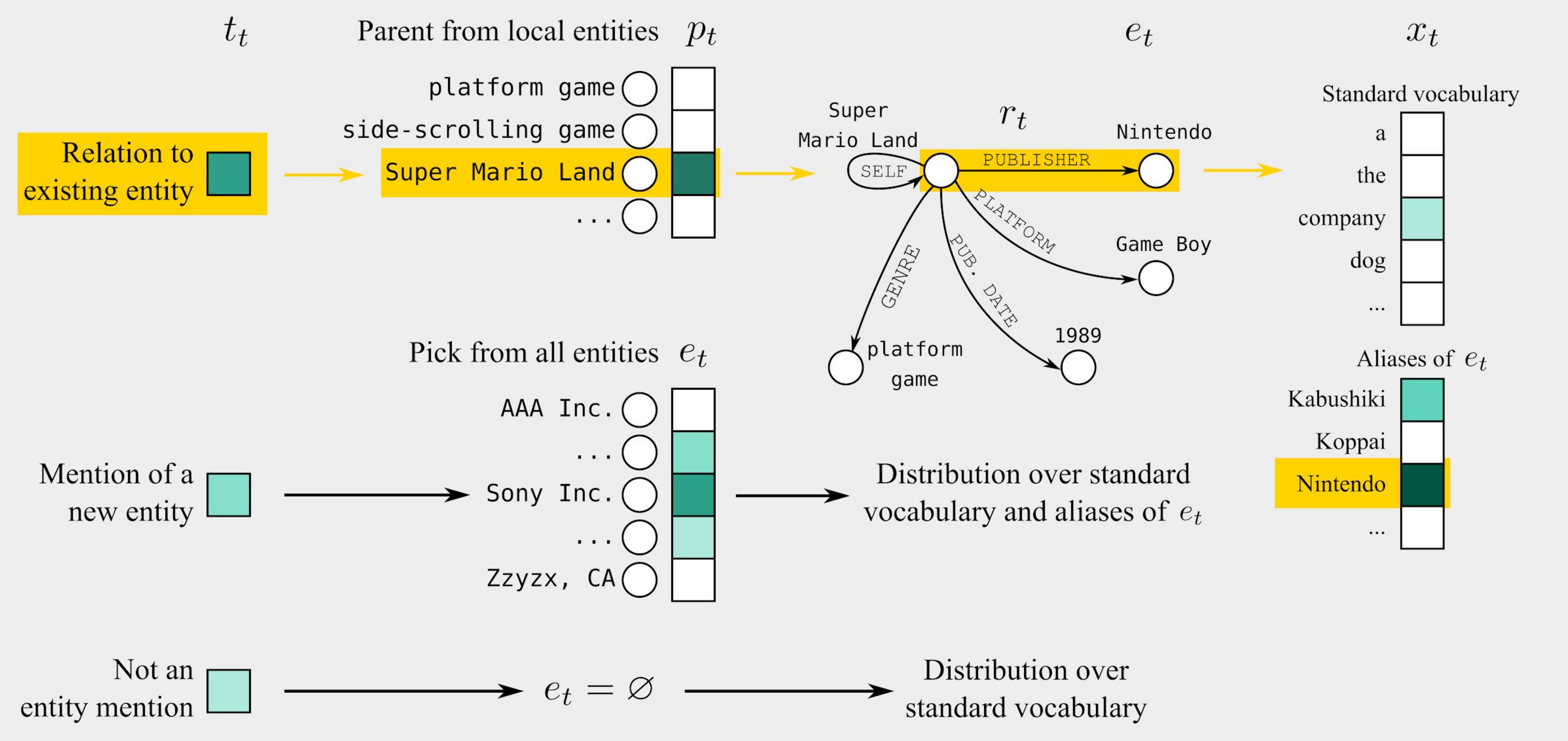
Motivating Example

[Super Mario Land] is a *[1989]* *[side-scrolling]* *[platform video game]* developed and publised by *[Nintendo]* as a *[launch title]* for their *[Game Boy]* *[handheld game console]*.



Generative Story & Model

Super Mario Land is a 1989 side-scrolling platform video game developed and published by [Nintendo]



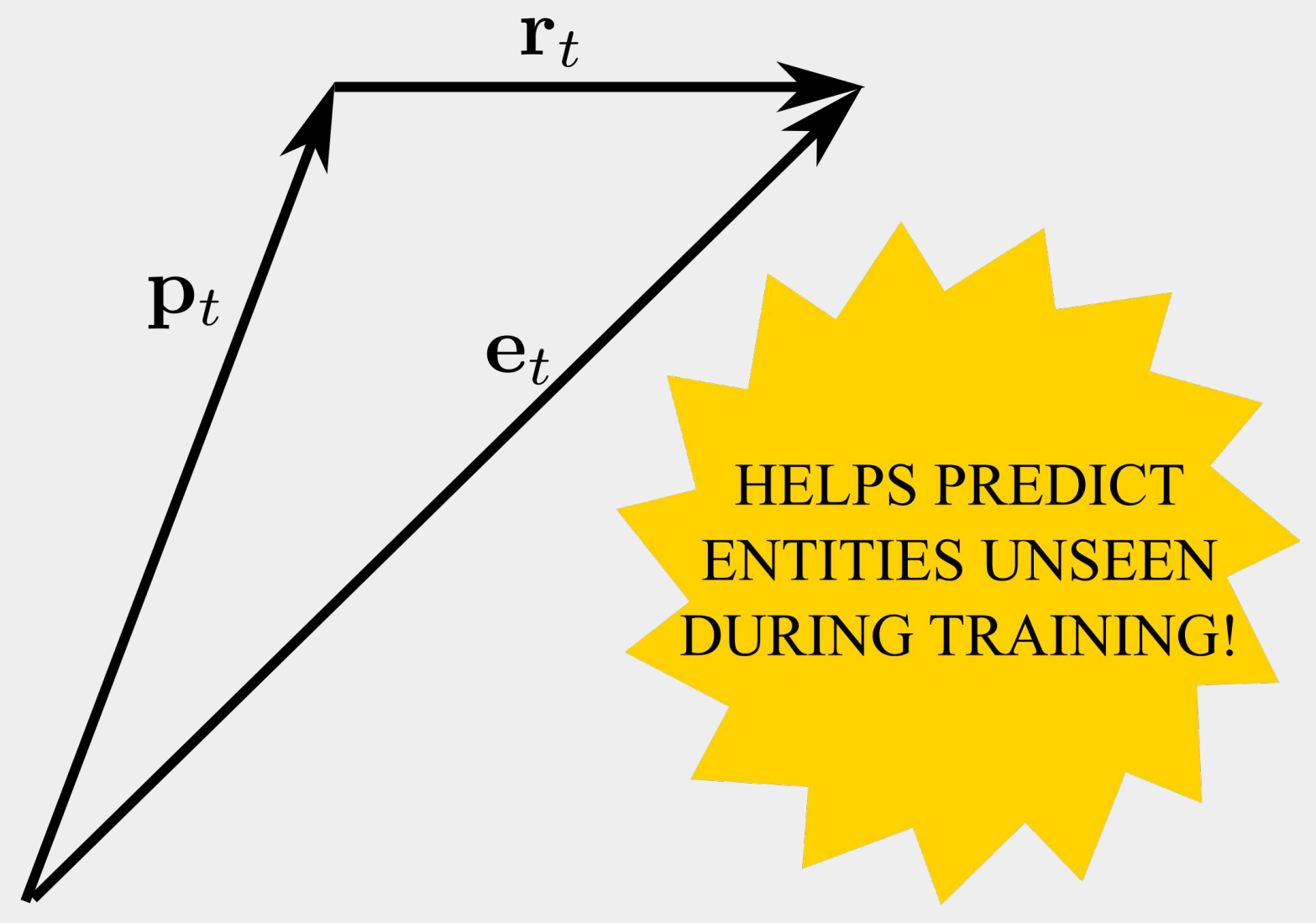
Inference

	Language model		KGLM
Problem	$p(X)$	$\neq$	$p(X, T, P, R, E)$

Importance Sampling

Solution $p(X) \approx \frac{1}{N}\sum_{T,P,R,E\sim q}\frac{p(X,T,P,R,E)}{q(T,P,R,E\mid X)}$

TransE



$$\mathcal{L}_{\text{TransE}} = \max\left(0, \gamma + \delta(\mathbf{p}_t + \mathbf{r}_t, \mathbf{e}_t) - \delta(\mathbf{p}_t^- + \mathbf{r}_t, \mathbf{e}_t^-)\right)$$

Linked WikiText-2 Dataset

Example Annotation

Tokens	Super Mario Land	is	a	1989	side - scrolling	platform video game
Mention Type	new			related	new	related
Entity Mentioned	SML			4-21-1989	SIDE_SCROLL	PVG
Relation				pub. date		genre
Parent Entity				SML		SML

developed	and	published	by	Nintendo	as	a	launch title	for	their	Game Boy
				related			new			related
				NIN			LT			GAME_BOY
				publisher						manuf./platform
				SML						NIN/SML

Dataset Statistics

	Train	Dev	Test
Documents	600	60	60
Tokens	2M	200K	236K
Vocabulary Size	33K	-	-
Mention Tokens	207K	21K	24K
Mention Spans	123K	12K	15K
Unique Entities	41K	5.4K	5.6K
Unique Relations	1.2K	484	504

Resources

Dataset

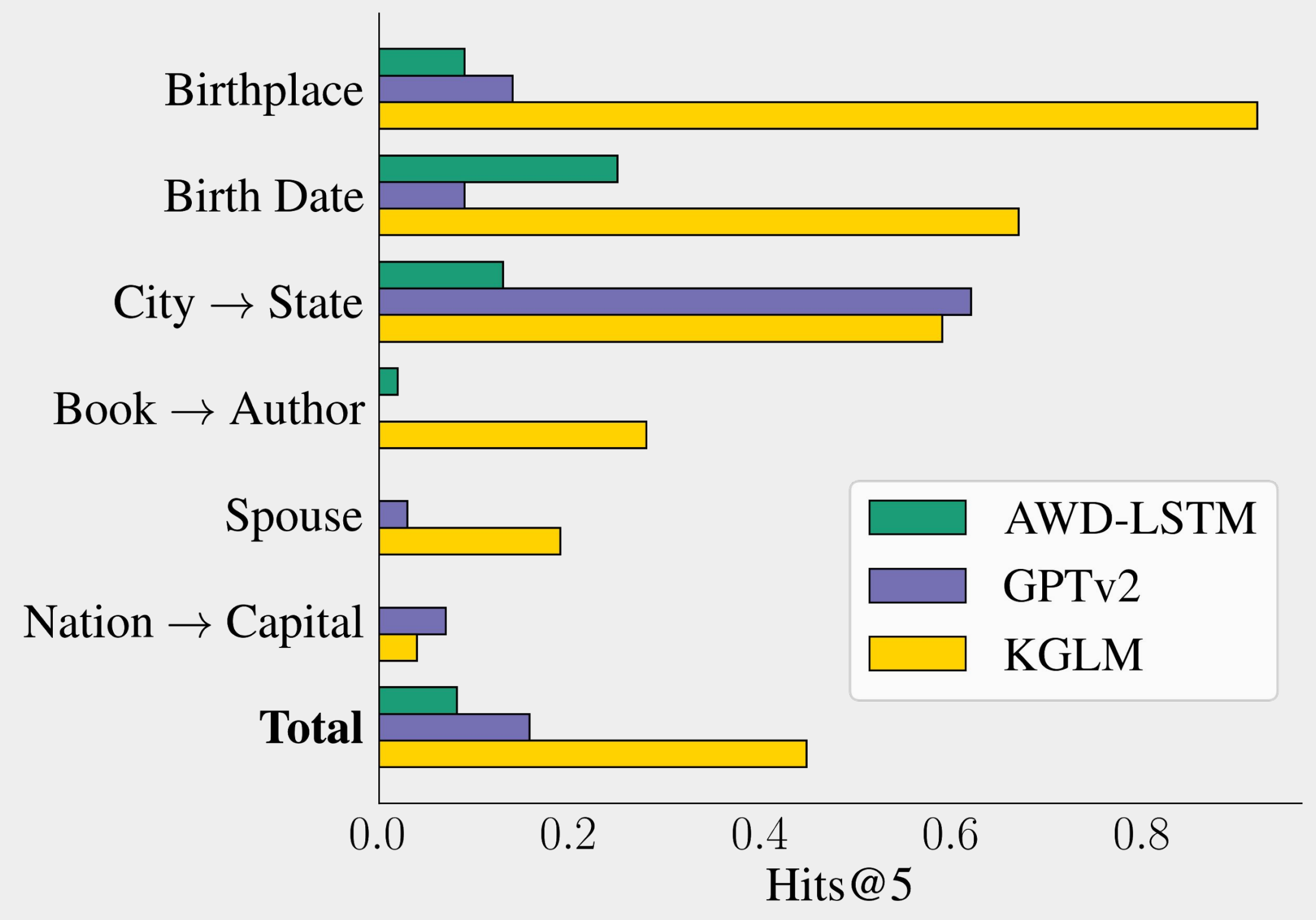
rloganiv.github.io/linked-wikitext-2

Code

github.com/rloganiv/kglm-model

Experiments

Fact Completion



Fact Completion Examples

	Input Sentence	Gold	GPT-2	KGLM
Both Correct	Paris Hilton was born in ____	New York City	New	New
	Arnold Schwarzenegger was born on ____	1947-07-30	July	30
KGLM Correct	Bob Dylan was born in ____	Duluth	New	Duluth
	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
	Kanye West is married to ____	Kim Kardashian	Kim	the
Both Wrong	The capital of India is ____	New Dehli	the	a
	Madonna is married to	Carlos Leon	a	Alex

Perplexity

	PPL	UPP
EntityNLM[*]	85.4	189.2
EntityCopyNet[*]	76.1	144.0
AWD-LSTM	74.8	165.8
KGLM[*]	**44.1**	**88.5**

[*]Obtained using importance sampling

Unknown Penalty

$$\mathrm{P}_{\mathrm{UPP}}(w_{\text{unk}}) = \frac{P(w_{\text{unk}})}{V_{\text{total}}/V_{\text{vocab}}}$$

For any questions, email: `rlogan@uci.edu`