Part 3: Knowledge-Augmented PreTraining for Reasoning

ACL 2023

Tutorial on Complex Reasoning in Natural Language

Michihiro Yasunaga

Stanford University

61 ACL 2023

Recap: Knowledge

Knowledge is available in various forms

Text

- Diverse & contextual knowledge

WIKIPEDIA

PubMed

Statue of Liberty

Article Talk

From Wikipedia, the free encyclopedia

For other uses, see Statue of Liberty (disamb

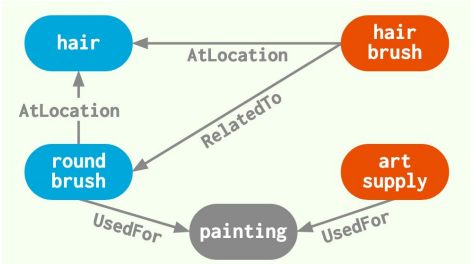
The **Statue of Liberty** (***Liberty Enlightening the World***; French: *La Liberté éclairant le monde*) is a colossal neoclassical sculpture on Liberty Island in New York Harbor in New York City, in the United States. The copper statue, a gift from the people of France, was designed by French sculptor Frédéric Auguste Bartholdi and its metal framework was built by Gustave Eiffel. The statue was dedicated on October

Knowledge Graph (KG)

- Structured background knowledge

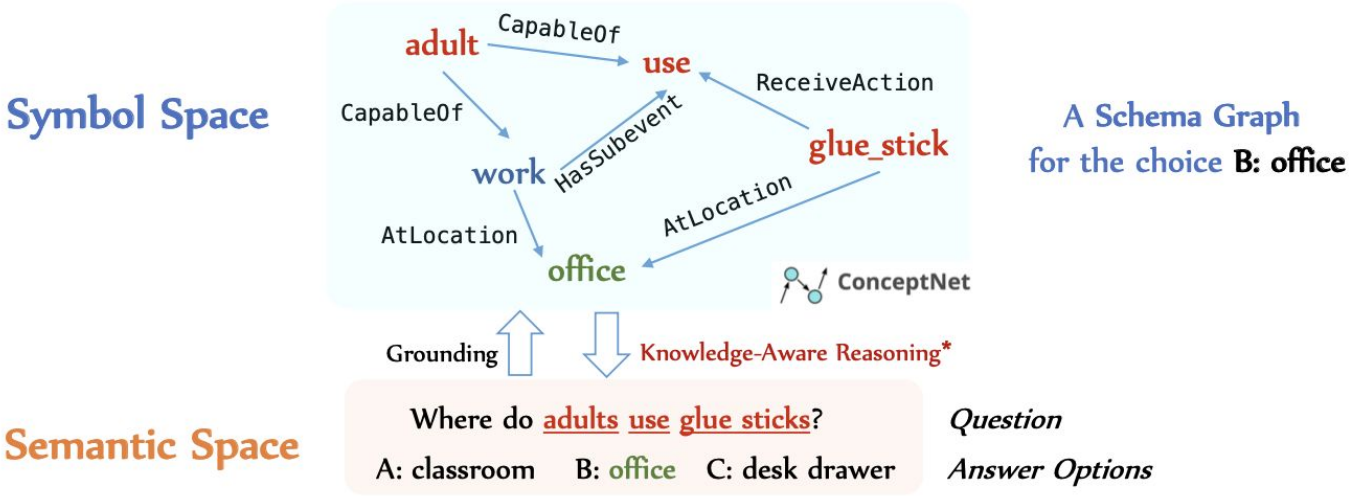
ConceptNet yago select knowledge

Freebase



Knowledge helps complex reasoning

Reasoning often involves combining multiple pieces of knowledge

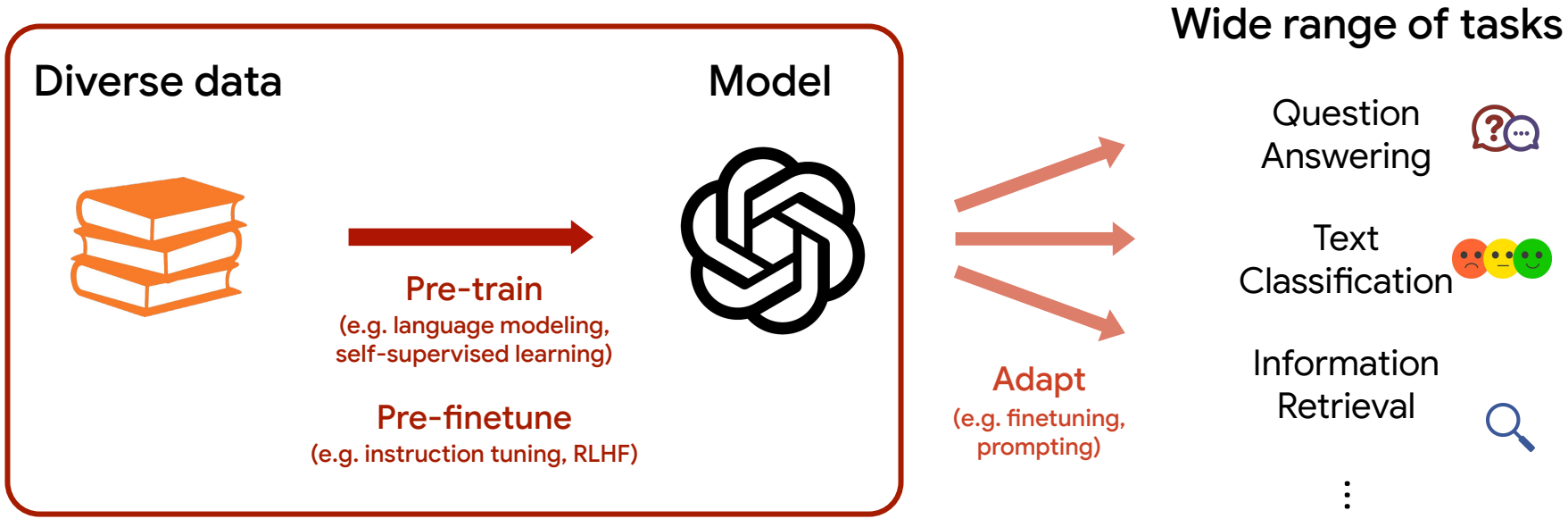


Lin+2019; Yasunaga+2021

This section:
Knowledge-augmented **Pre-**Training

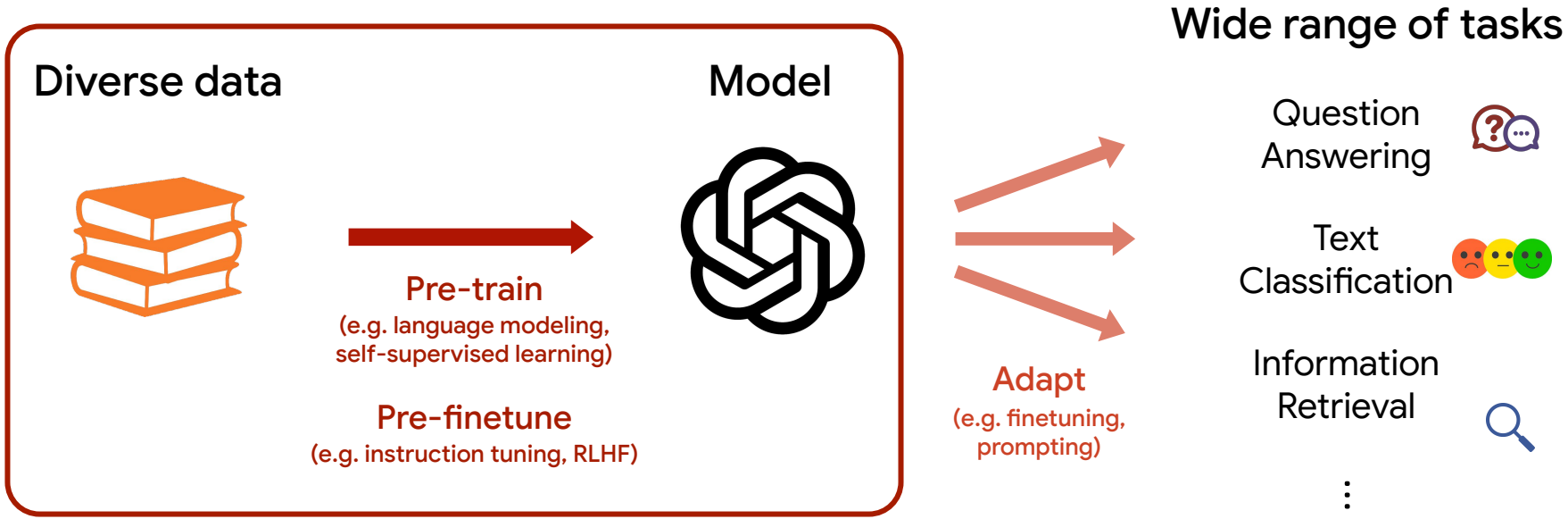
What is Pre-Training (and Pre-Finetuning)

- **Key:** learn from diverse data (e.g. through self-supervised learning)



Why Pre-Training (and Pre-Finetuning)?

- Help a broad range of downstream tasks
- Make adaptation efficient (e.g. few-shot finetuning/prompting)

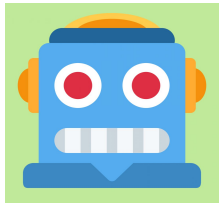


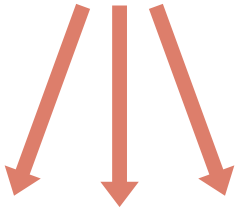
Goal: Knowledge-augmented Pre-Training

Text

- Diverse & contextual knowledge

Knowledge-augmented Pre-training





Knowledge- & Reasoning-intensive Tasks

Knowledge Graph (KG)

- Structured background knowledge

Outline of Knowledge-augmented Pre-training

Integrate textual knowledge		REALM [ICML 2020] CDLM [EMNLP 2021] LinkBERT [ACL 2022]
Integrate structured knowledge	Knowledge graph as training objective	WKLM [ICLR 2020] KEPLER [TACL 2021] JAKET [AAAI 2022]
	Knowledge graph as input context	ERNIE [ACL 2019] CoLAKE [COLING 2020] DRAGON [NeurIPS 2022]

Integrate Textual Knowledge

Integrate textual knowledge		REALM [ICML 2020] CDLM [EMNLP 2021] LinkBERT [ACL 2022]
Integrate structured knowledge	Knowledge graph as training objective	WKLM [ICLR 2020] KEPLER [TACL 2021] JAKET [AAAI 2022]
	Knowledge graph as input context	ERNIE [ACL 2019] CoLAKE [COLING 2020] DRAGON [NeurIPS 2022]

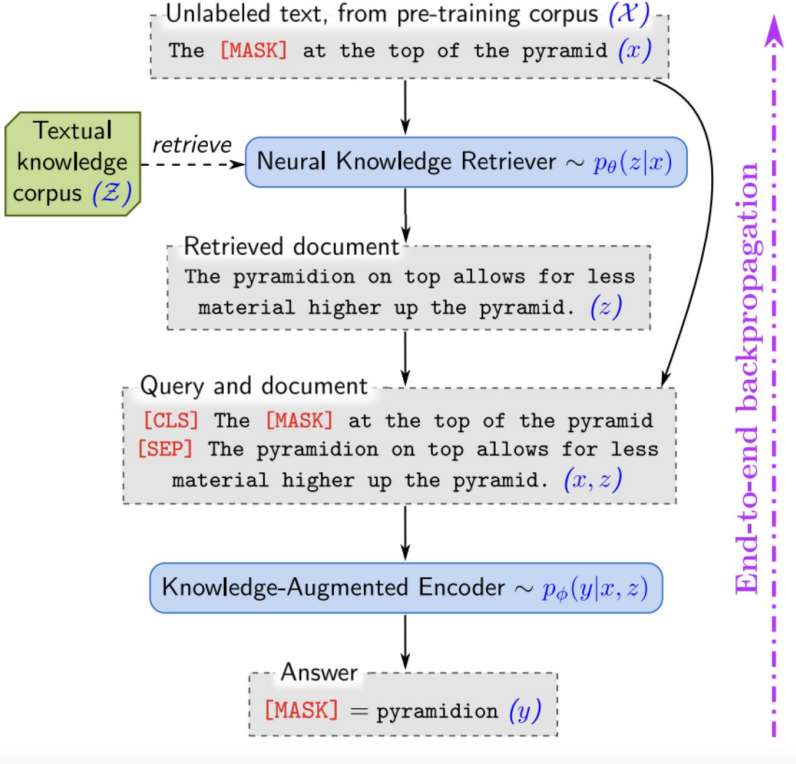
REALM: Retrieval-Augmented Language Model Pre-Training

Method

- Given input text (e.g. with masked tokens), **retrieve** relevant documents from a knowledge corpus to answer it

Key

- DPR (from previous section) was retrieval-augmented finetuning for QA. REALM is a **self-supervised pre-training** version.



Guu et al. "Realm: Retrieval-augmented language model pre-training." ICML 2020.

REALM: Retrieval-Augmented Language Model Pre-Training

Result

- Improve knowledge-intensive NLP (e.g. Open-domain QA)

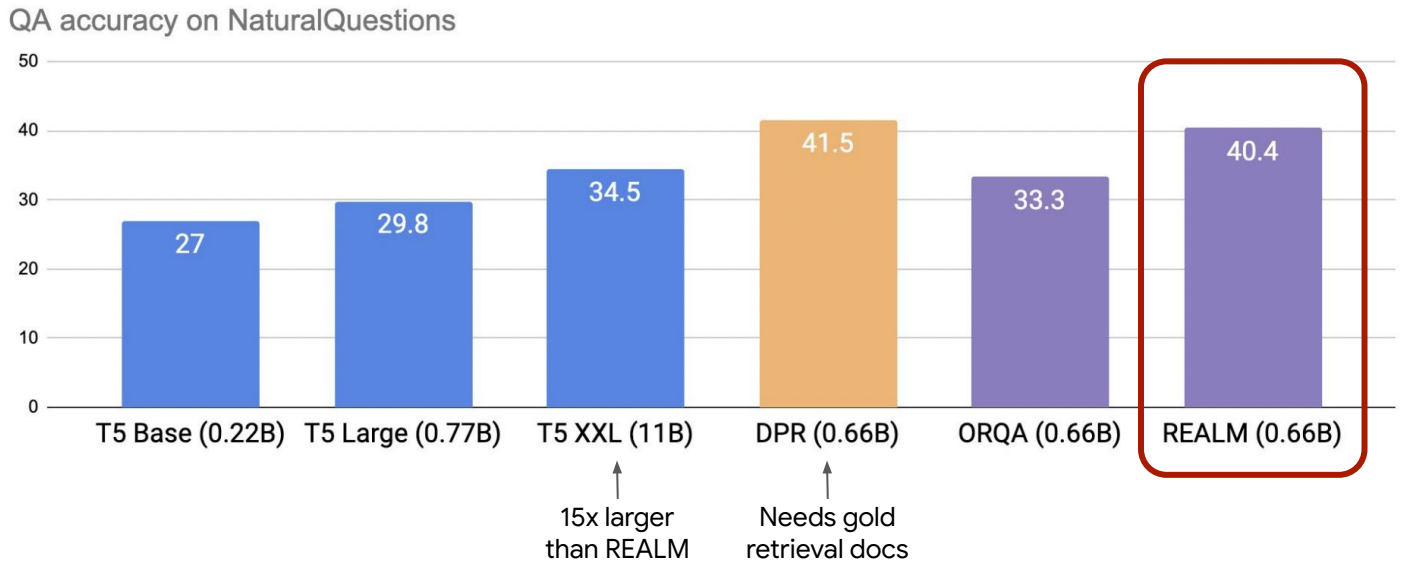


Figure from https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1224/slides/cs224n-2022-lecture15-guu.pdf

CDLM: Cross-Document Language Modeling

Method

- Retrieve related docs (e.g. same topic) and pre-train LM on concatenated context
- Internalize knowledge during pretraining. Retrieval is optional at test time.

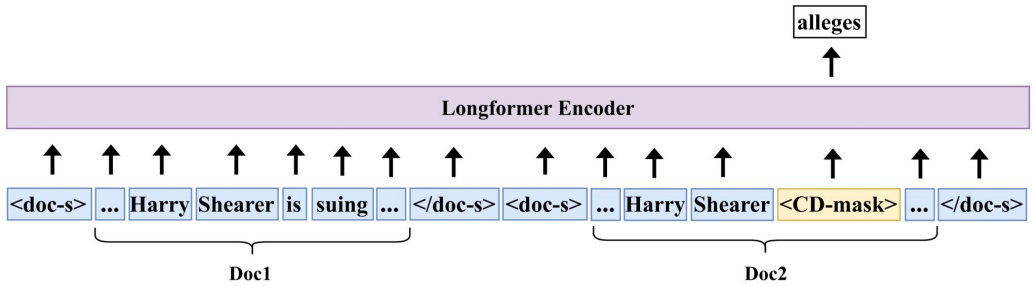
Collection of related documents

Doc 1: "*Harry Shearer* is *suing* *Vivendi's* Universal Music for $125 million for allegedly fraudulent ..."

Doc 2:"...*Harry Shearer* *alleges* parent company of Universal Music and StudioCanal withheld millions..."

Doc 3: "*Shearer* was then *joined in the lawsuit* with *StudioCanal* and its French parent *Vivendi* by his co-stars"

LM pretraining



Caciularu et al. "CDLM: Cross-Document Language Modeling." EMNLP 2021.

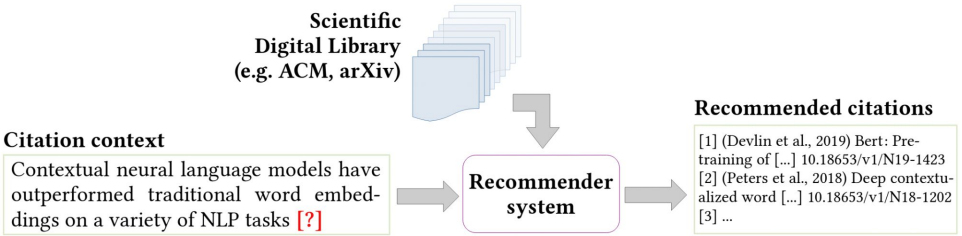
CDLM: Cross-Document Language Modeling

Result

- Improve cross-document NLP (e.g. citation recommendation, coreference resolution)

Takeaway

- Retrieval-augmented pre-training helps **cross-document reasoning**



(Citation recommendation illustration: https://arxiv.org/pdf/2108.07571.pdf)

Model	AAN	OC	S2orc	PAN
SMASH (2019)[5]	80.8	-	-	-
SMITH (2020)[5]	85.4	-	-	-
BERT-HAN (2020)	65.0	86.3	90.8	**87.4**
GRU-HAN+CDA (2020)	75.1	89.9	91.6	78.2
BERT-HAN+CDA (2020)	82.1	87.8	92.1	86.2
Longformer	85.4	93.4	95.8	80.4
Local CDLM	83.8	92.1	94.5	80.9
Rand CDLM	85.7	93.5	94.6	79.4
Prefix CDLM	87.3	94.8	94.7	81.7
CDLM	**88.8**	**95.3**	**96.5**	82.9

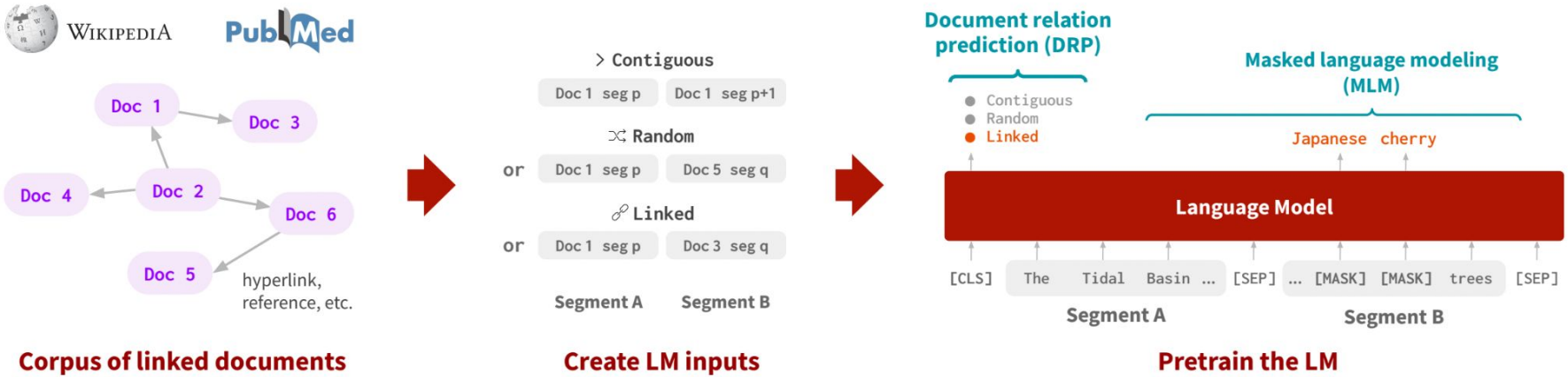
Table 4: F_1 scores over the document matching benchmarks' test sets.

Caciularu et al. "CDLM: Cross-Document Language Modeling." EMNLP 2021.

LinkBERT: Pretraining Language Models with Document Links

Method

- Retrieve related docs (e.g. hyperlinked) and pre-train LM on concatenated context
- Add auxiliary objective to learn doc relations
- Internalize knowledge during pretraining. Retrieval is optional at test time.

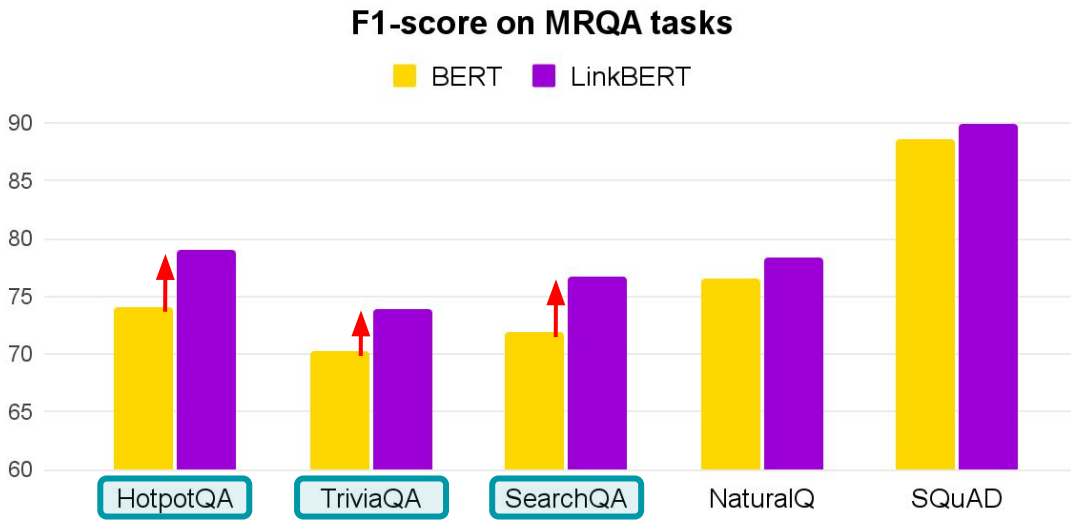


Yasunaga et al. "LinkBERT: Pretraining Language Models with Document Links". ACL 2022.

LinkBERT: Pretraining Language Models with Document Links

Result

- Improve knowledge-intensive NLP
- Improve multi-hop / multi-document reasoning



Yasunaga et al. "LinkBERT: Pretraining Language Models with Document Links". ACL 2022.

LinkBERT: Pretraining Language Models with Document Links

Takeaway

- Retrieval-augmented pre-training (⇒ multi-document in context) helps learn **multi-hop reasoning**

HotpotQA example

Question: **Roden Brothers** were taken over in 1953 by a group headquartered in which Canadian city?

Doc A: **Roden Brothers** was founded June 1, 1891 in Toronto, Ontario, Canada by Thomas and Frank Roden. In the 1910s the firm became known as Roden Bros. Ltd. and were later taken over by **Henry Birks and Sons** in 1953. ...

Doc B: **Birks Group** (formerly Birks & Mayors) is a designer, manufacturer and retailer of jewellery, timepieces, silverware and gifts ... The company is headquartered in **Montreal**, Quebec, ...

LinkBERT predicts: "Montreal" (✓) **BERT predicts: "Toronto" (✗)**

Yasunaga et al. "LinkBERT: Pretraining Language Models with Document Links". ACL 2022.

Summary so far

Integrate textual knowledge		REALM - retrieve relevant docs CDLM - learn relevant docs LinkBERT - learn relevant docs & doc relations
Integrate structured knowledge	Knowledge graph as training objective	WKLM - KG entity objective KEPLER - KG link objective JAKET - both entity and link objectives
	Knowledge graph as input context	ERNIE - contextualize entity emb CoLAKE - contextualize KG triplet DRAGON - contextualize KG subgraph

Integrate Knowledge Graph as Training Objective

Integrate textual knowledge		REALM [ICML 2020] CDLM [EMNLP 2021] LinkBERT [ACL 2022]
Integrate structured knowledge	Knowledge graph as training objective	WKLM [ICLR 2020] KEPLER [TACL 2021] JAKET [AAAI 2022]
	Knowledge graph as input context	ERNIE [ACL 2019] CoLAKE [COLING 2020] DRAGON [NeurIPS 2022]

Convenient (KG not needed at test time)

Expressive model

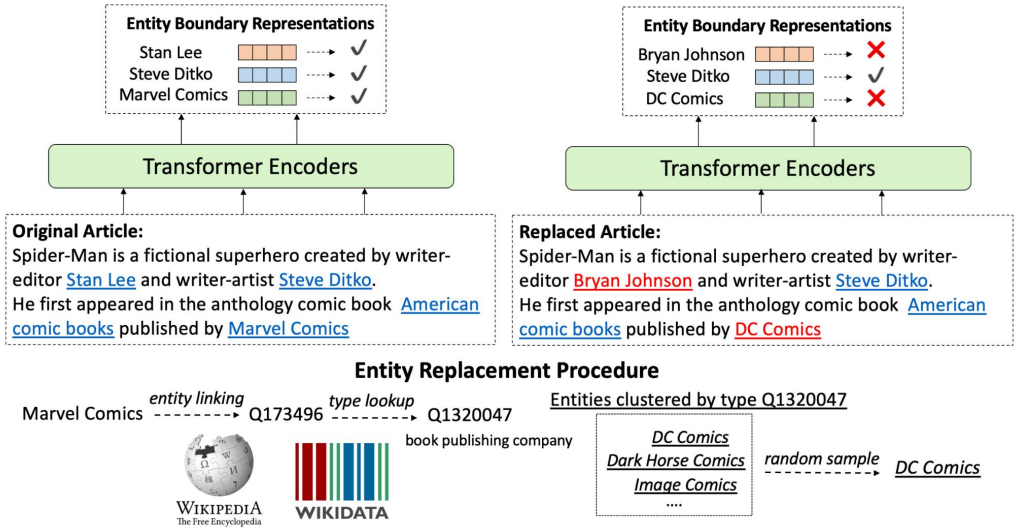
WKLM: Weakly Supervised Knowledge-Pretrained Language Model

Idea

- Besides language modeling, teach entity knowledge from KG

Method

- Replace entity mentions in text by false entities
- Add **auxiliary objective** to **learn true/false entities**



Xiong, et al. "Weakly Supervised Knowledge-Pretrained Language Model." ICLR 2020.

WKLM: Weakly Supervised Knowledge-Pretrained Language Model

Result

- Improve knowledge-intensive NLP (e.g. QA, entity typing)

Takeaway

- KG can augment the pre-training objective/task for LM

Model	SQuAD (F1)	TriviaQA (F1)	Quasar-T (F1)	FIGER (acc)
WKLM	91.3	56.7	49.9	60.21
WKLM w/o MLM	87.6	52.5	48.1	58.44
BERT + 1M Updates	91.1	56.3	48.2	54.17

Much worse without MLM

Much worse training for longer, compared to using the entity replacement loss

Xiong, et al. "Weakly Supervised Knowledge-Pretrained Language Model." ICLR 2020.

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

Motivation

- KG link prediction ⇒ teach how to reason about entity relations

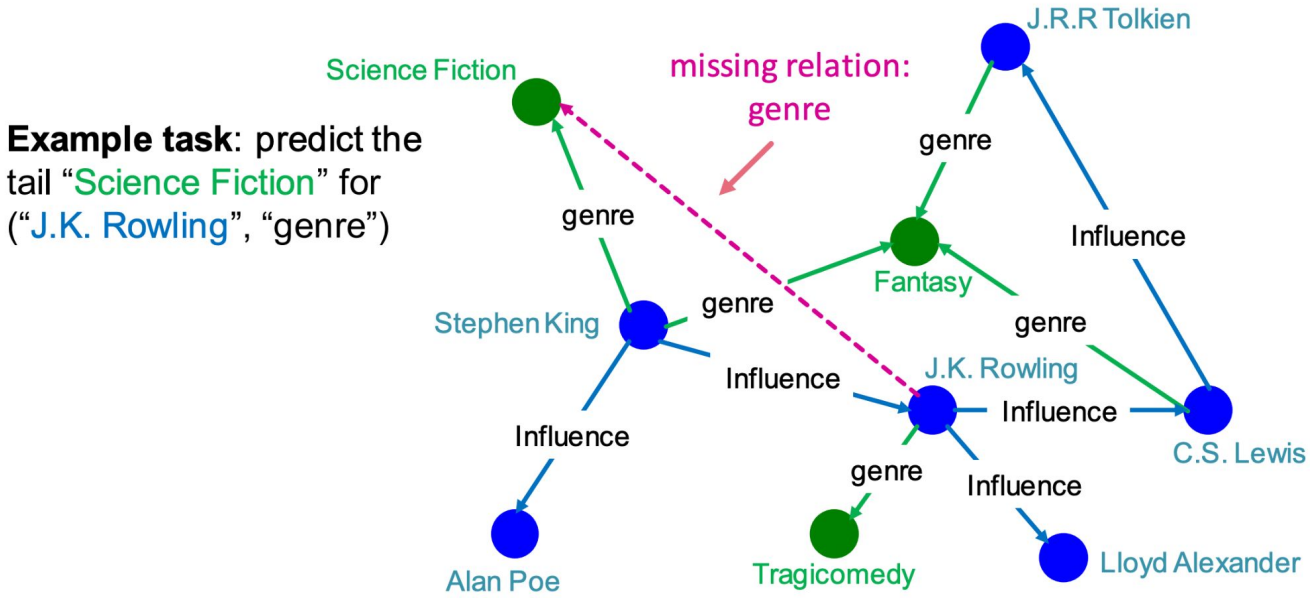
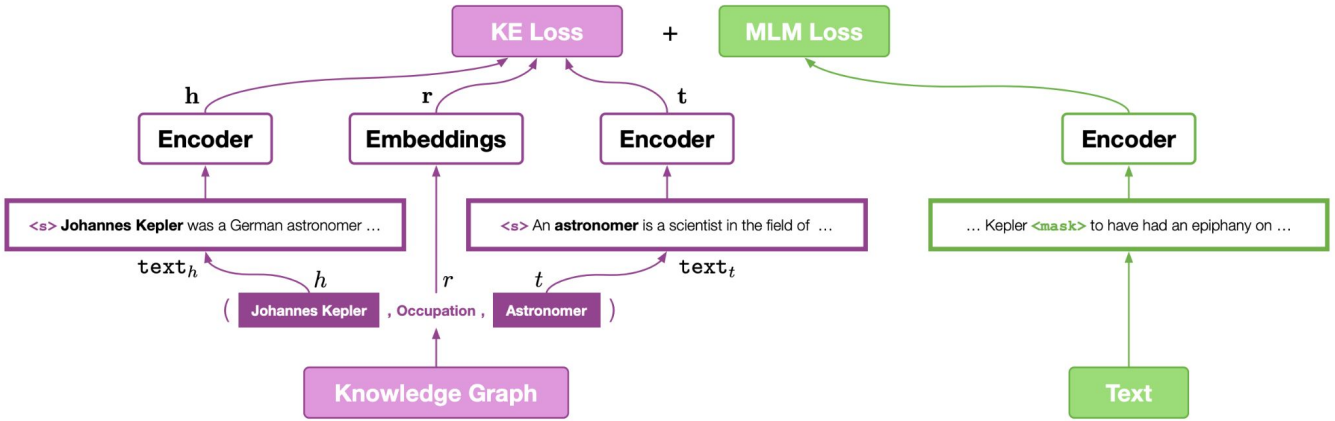


Figure from http://web.stanford.edu/class/cs224w/slides/10-kg.pdf

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

Method

- Add KG **link prediction objective** in LM pre-training
  - Predict whether (head, relation, tail) forms a link or not
  - TransE head: || **h** + **r** – **t** ||



Wang, et al. "KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation". TACL 2021.

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

Result

- Improve knowledge-intensive NLP and KG link prediction

Takeaway

- Besides entities, KG **structure (links)** can augment LM pre-training objective

Model	P	R	F-1
$\text{ERNIE}_{\text{BERT}}$	70.0	66.1	68.0
$\text{KnowBert}_{\text{BERT}}$	**73.5**	64.1	68.5
RoBERTa	70.4	71.1	70.7
$\text{ERNIE}_{\text{RoBERTa}}$	**73.5**	68.0	70.7
$\text{KnowBert}_{\text{RoBERTa}}$	71.9	69.9	70.9
KEPLER-Wiki	71.5	**72.5**	**72.0**

Table 5: Precision, recall and F-1 on TACRED (%).

Model	MR	MRR	HITS@1	HITS@3	HITS@10
DKRL (Xie et al., 2016)	78	23.1	5.9	32.0	54.6
RoBERTa	723	7.4	0.7	1.0	19.6
KEPLER-Wiki	32	35.1	15.4	46.9	71.9
KEPLER-Cond	**28**	**40.2**	**22.2**	**51.4**	**73.0**

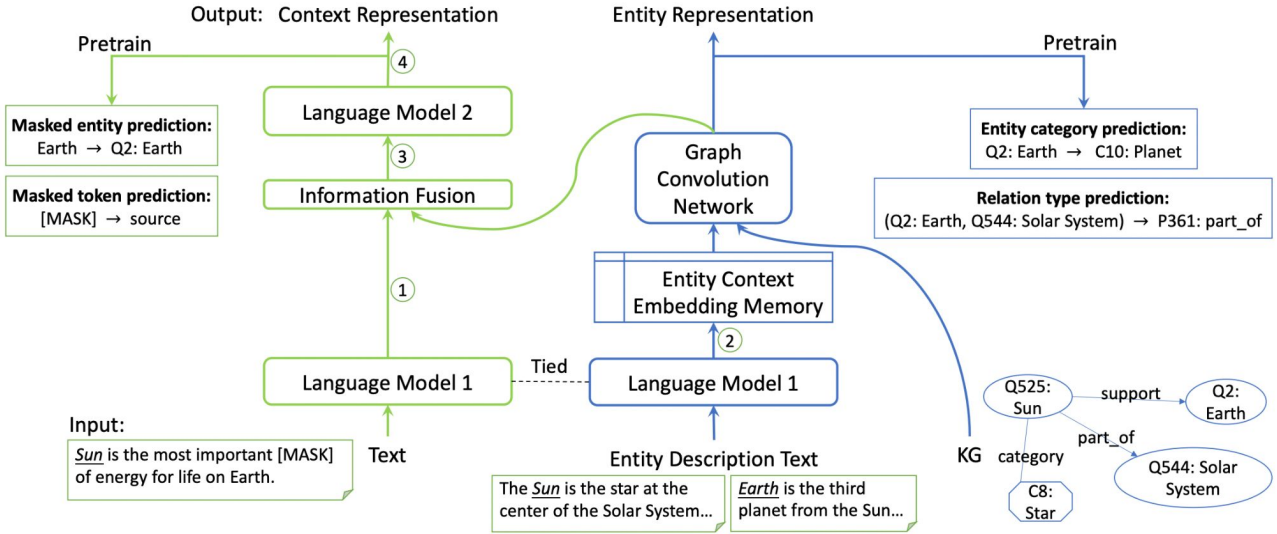
(b) Inductive results on Wikidata5M (% except MR).

Wang, et al. "KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation". TACL 2021.

JAKET: Joint Pre-training of Knowledge Graph and Language Understanding

Method

- Add **both KG entity and link** prediction objectives in LM training



Yu, Zhu, et al. "JAKET: Joint Pre-training of Knowledge Graph and Language Understanding". AAAI 2022.

JAKET: Joint Pre-training of Knowledge Graph and Language Understanding

Result

- Improve multi-hop QA

Takeaway

- KG-augmented objective can help **multi-hop reasoning**

MetaQA example (2-hop): *"Who acted in the movies directed by Erik Poppe?"*

Model	KG-Full		KG-50%	
	1-hop	2-hop	1-hop	2-hop
RoBERTa	90.2	70.8	61.5	39.3
RoB+G+M	91.4	72.6	62.5	40.8
JAKET	**93.9**	**73.2**	**63.1**	**41.9**

Table 2: Results on the MetaQA dataset over 1-hop and 2-hop questions under *KG-Full* and *KG-50%* settings. RoB+G+M is the abbreviation for the baseline model RoBERTa+GNN+M.

Yu, Zhu, et al. "JAKET: Joint Pre-training of Knowledge Graph and Language Understanding". AAAI 2022.

Summary so far

Integrate textual knowledge		REALM - retrieve relevant docs CDLM - learn relevant docs LinkBERT - learn relevant docs & doc relations
Integrate structured knowledge	Knowledge graph as training objective	WKLM - KG entity objective KEPLER - KG link objective JAKET - both entity and link objectives
	Knowledge graph as input context	ERNIE - contextualize entity emb CoLAKE - contextualize KG triplet DRAGON - contextualize KG subgraph

Integrate Knowledge Graph as Input Context

Integrate textual knowledge		REALM [ICML 2020] CDLM [EMNLP 2021] LinkBERT [ACL 2022]
Integrate structured knowledge	Knowledge graph as training objective	WKLM [ICLR 2020] KEPLER [TACL 2021] JAKET [AAAI 2022]
	Knowledge graph as input context	ERNIE [ACL 2019] CoLAKE [COLING 2020] DRAGON [NeurIPS 2022]

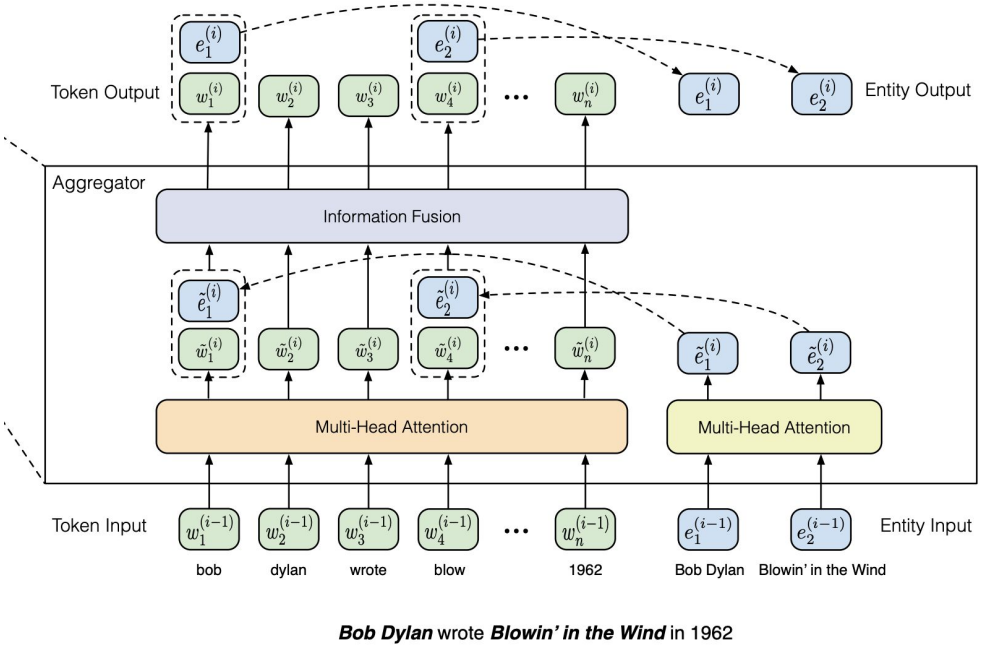
Convenient (KG not needed at test time)

Expressive model

ERNIE: Enhanced Language Representation with Informative Entities

Method

- Add **entity embeddings** in the LM context
- Entity emb: TransE from Wikidata
- Entity linking: TAGME



Zhang, Zhengyan, et al. "ERNIE: Enhanced Language Representation with Informative Entities." ACL 2019.

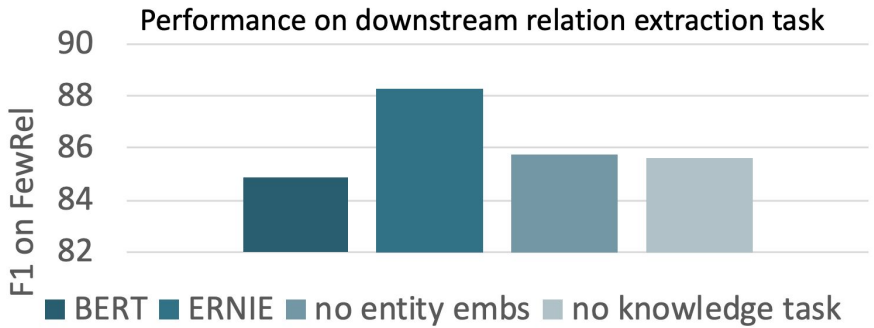
ERNIE: Enhanced Language Representation with Informative Entities

Result

- Improve knowledge-intensive NLP (e.g., relation extraction, QA)

Takeaway

- Pioneered the idea of integrating **KG information as input** in LM pre-training

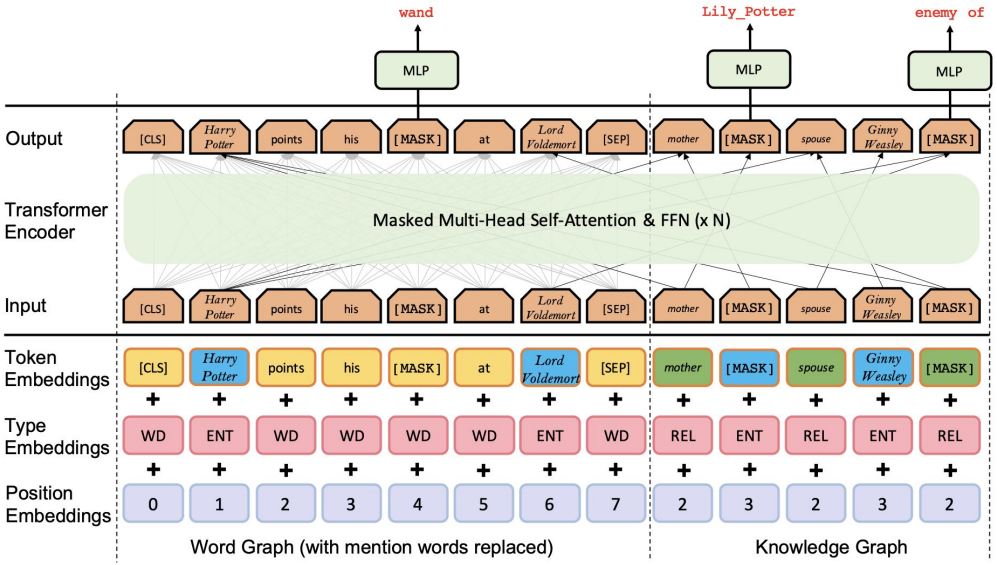


Zhang, Zhengyan, et al. "ERNIE: Enhanced Language Representation with Informative Entities." ACL 2019.

CoLAKE: Contextualized Language and Knowledge Embedding

Method

- Add **KG triplets** in the LM context
  (⇒ bring neighbor entities and relations in context, besides direct entity mentions in text)
- Masked token prediction for both text and KG sides



Text
Harry Potter points his wand at Lord Voldemort

KG triplets
(Harry Potter, mother, Lily Potter),
(Harry Potter, spouse, Ginny Weasley),
(Harry Potter, enemy of, Lord Voldemort)

Sun, et al. "CoLAKE: Contextualized Language and Knowledge Embedding." COLING 2020

CoLAKE: Contextualized Language and Knowledge Embedding

Result

- Improve knowledge-intensive NLP and KG link prediction

Takeaway

- KG triplets bring in background knowledge and help **reason about related entities**

Model	MR ↓	MRR
	Transductive setting	
TransE (Bordes et al., 2013)	15.97	67.30
DistMult (Yang et al., 2015)	27.09	60.56
ComplEx (Trouillon et al., 2016)	26.73	61.09
RotatE (Sun et al., 2019)	30.36	70.90
CoLAKE	2.03	82.48
	Inductive setting	
DKRL (Xie et al., 2016)	168.21	8.18
CoLAKE	31.01	28.10

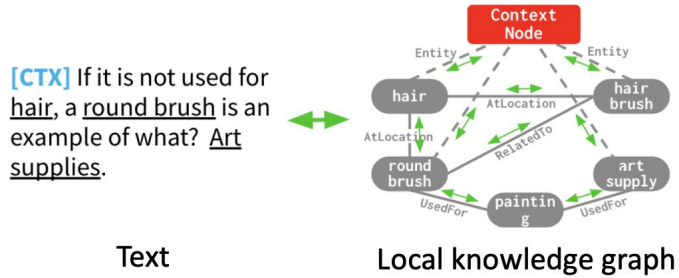
Table 5: The experimental results on word-knowledge graph completion task.

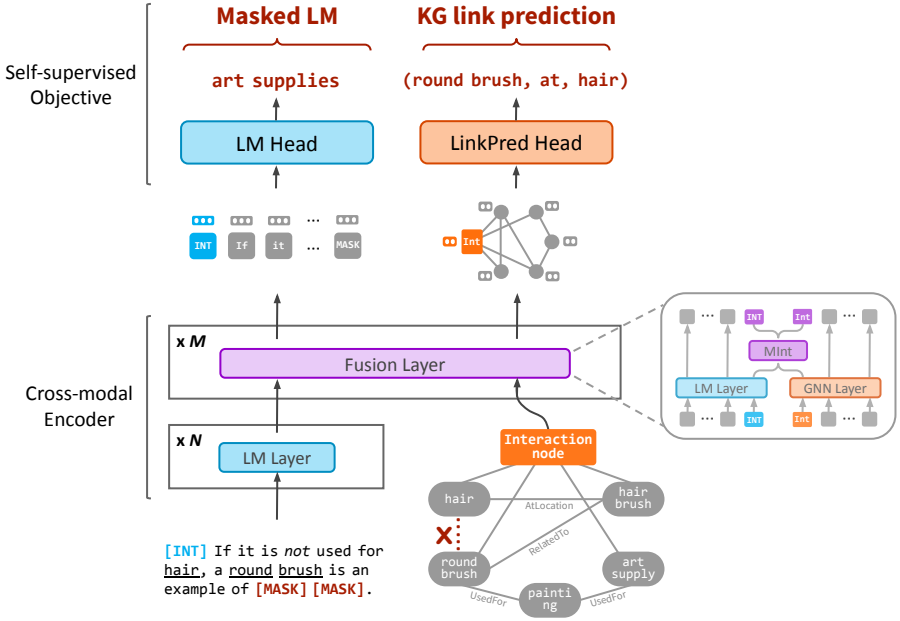
Sun, et al. "CoLAKE: Contextualized Language and Knowledge Embedding." COLING 2020

DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining

Method

- Add **KG subgraph** in the input context
- Text is contextualized by LM and KG is contextualized by GNN. The two are then contextualized bidirectionally



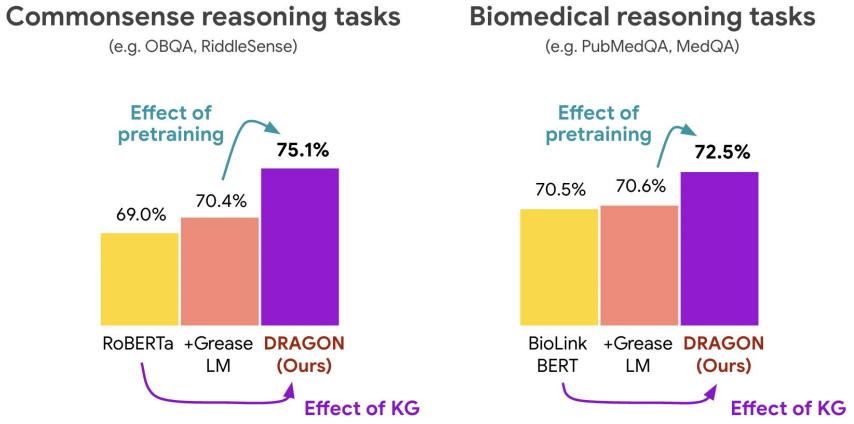


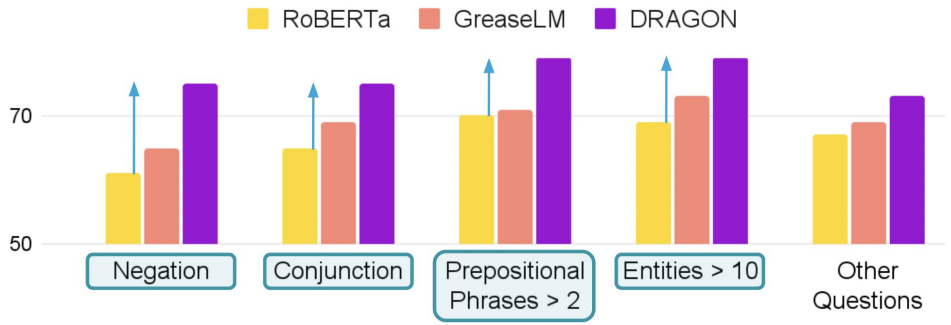
Yasunaga, et al. "DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining" NeurIPS 2022

DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining

Result

- Improve broad reasoning tasks (QA, commonsense, link prediction)
- Improve **complex reasoning (multi-hop, logical)**



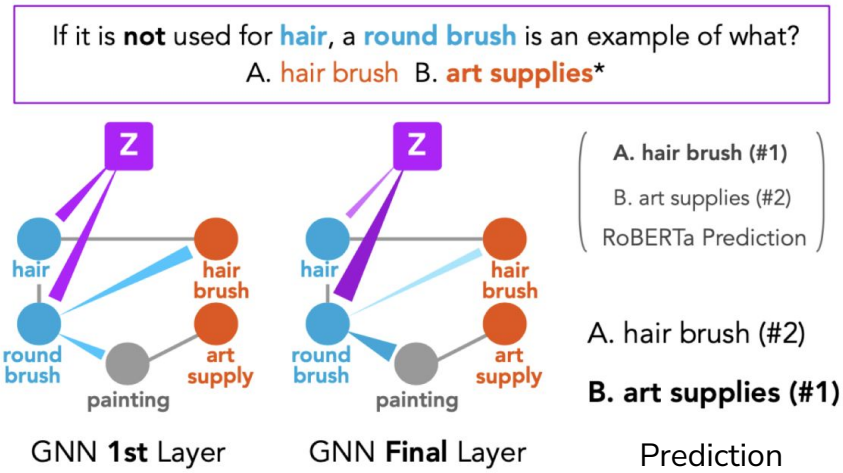


Yasunaga, et al. "DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining" NeurIPS 2022

DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining

Takeaway

- KG **graph structure** provides LM with a **scaffold** to perform complex reasoning about entities



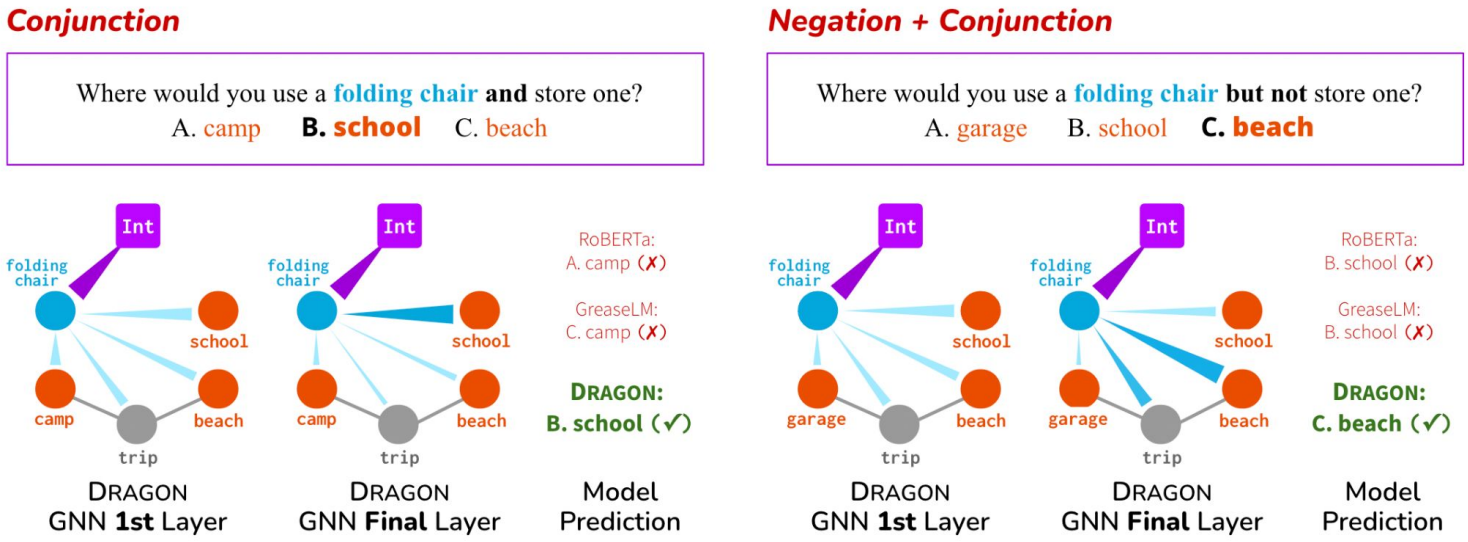
After several layers of fusion,
attention weight from text over **hair** decreases,
but attention weight over **round brush** and **painting** increases,
adjusting for the negation in text

Yasunaga, et al. "DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining" NeurIPS 2022

DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining

Takeaway

- KG **graph structure** provides LM with a **scaffold** to perform complex reasoning about entities



Yasunaga, et al. "DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining" NeurIPS 2022

Summary so far

Integrate textual knowledge		REALM - retrieve relevant docs CDLM - learn relevant docs LinkBERT - learn relevant docs & doc relations
Integrate structured knowledge	Knowledge graph as training objective	WKLM - KG entity objective KEPLER - KG link objective JAKET - both entity and link objectives
	Knowledge graph as input context	ERNIE - contextualize entity emb CoLAKE - contextualize KG triplet DRAGON - contextualize KG subgraph

Summary

Integrate textual knowledge		REALM - retrieve relevant docs CDLM - learn relevant docs LinkBERT - learn relevant docs & doc relations
Integrate structured knowledge	Knowledge graph as training objective	WKLM - KG entity objective KEPLER - KG link objective JAKET - both entity and link objectives
	Knowledge graph as input context	ERNIE - contextualize entity emb CoLAKE - contextualize KG triplet DRAGON - contextualize KG subgraph

Conclusion

Takeaways

- Knowledge can be integrated into LM in **self-supervised** ways
- Help a wide range of reasoning tasks

Open questions

- Can we integrate knowledge in **pre-finetuning** (e.g. instruction tuning, RLHF)?
- How can we build a **unified** model with all various knowledge sources?
- How can we ensure the models use and reason about knowledge **faithfully**?

Thak you!

https://cs.stanford.edu/~myasu/

@michiyasunaga

END