Empowering Language Models with Graph Learning

Michihiro Yasunaga

Joint work with Antoine Bosselut, Hongyu Ren, Xikun Zhang, Chris Manning, Percy Liang, Jure Leskovec

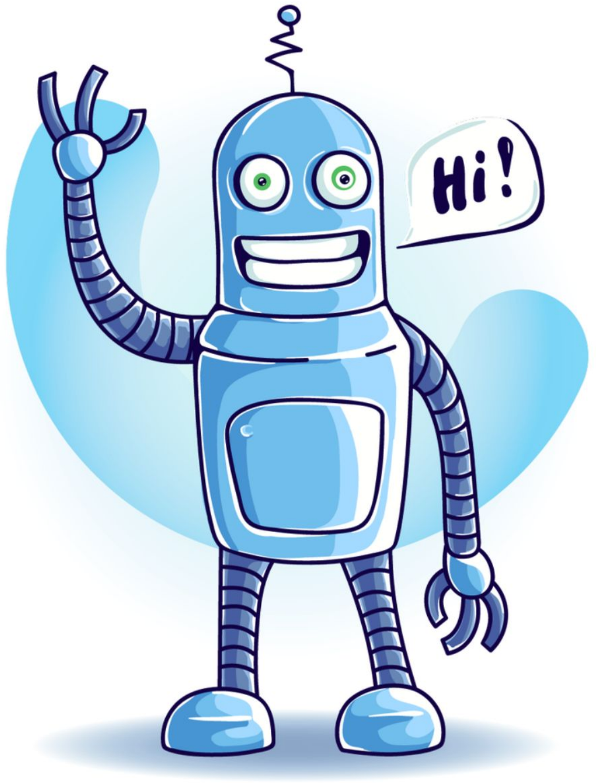
      

Stanford NLP

What is Natural Language Processing (NLP)?

- Automated understanding of natural language input

- Coherent generation of natural language output



NLP Applications

Machine Translation:

Question Answering:

bing

Personal Assistants:

Specialized Applications:

Specifio

NUANCE

LUMINOSO

Legal Documents

Health Records

Business Intelligence

Customer Research

Modern NLP – Powered by large language models


THE NEW YORKER

The Next Word

Where will predictive text take us?

Text by John Seabrook

The New York Times

A Breakthrough for A.I. Technology: Passing an 8th-Grade Science Test

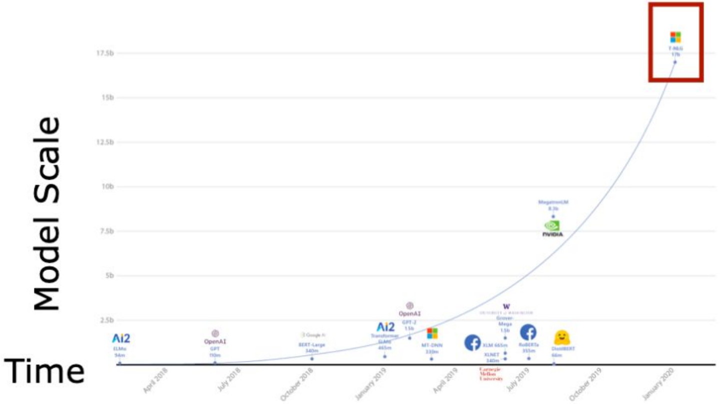
Vox

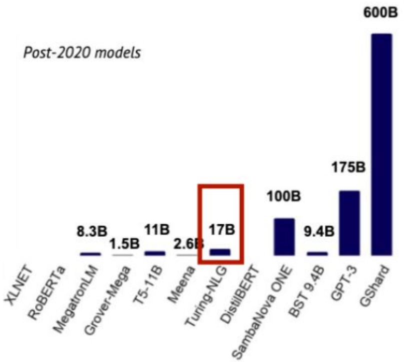
How I'm using AI to write my next novel

The New York Times

Meet GPT-3. It Has Learned to Code (and Blog and Argue).

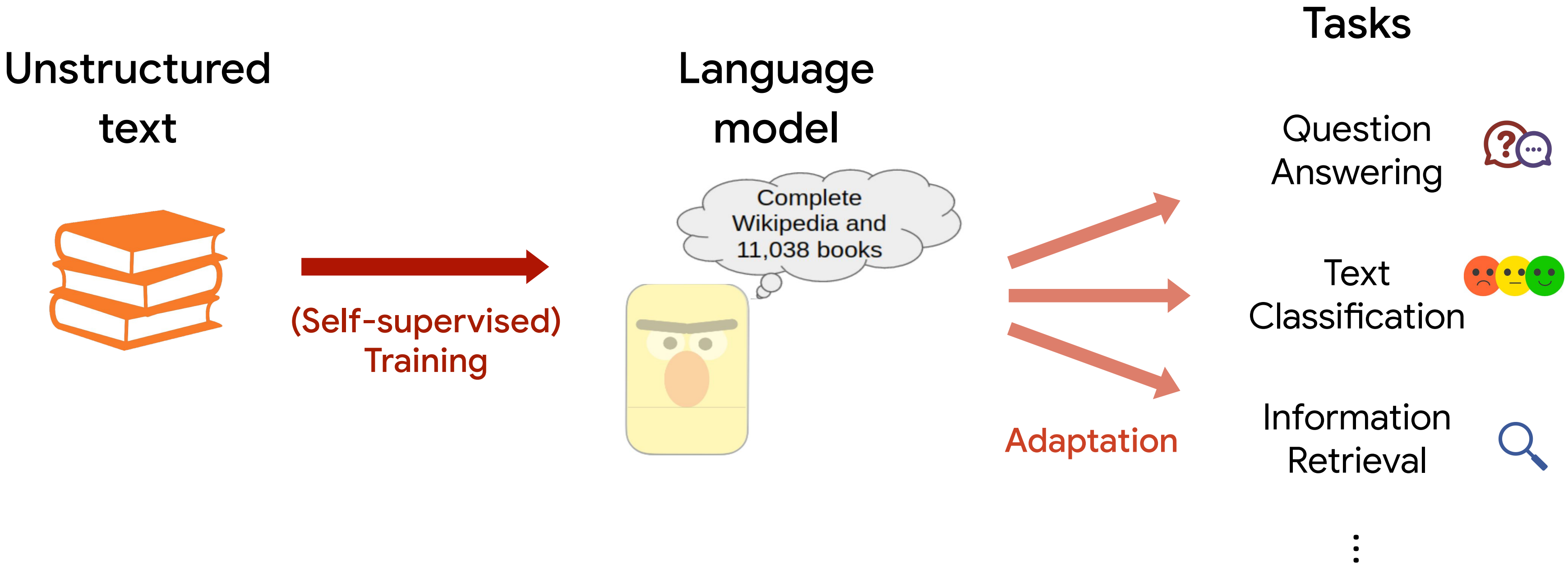
The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.





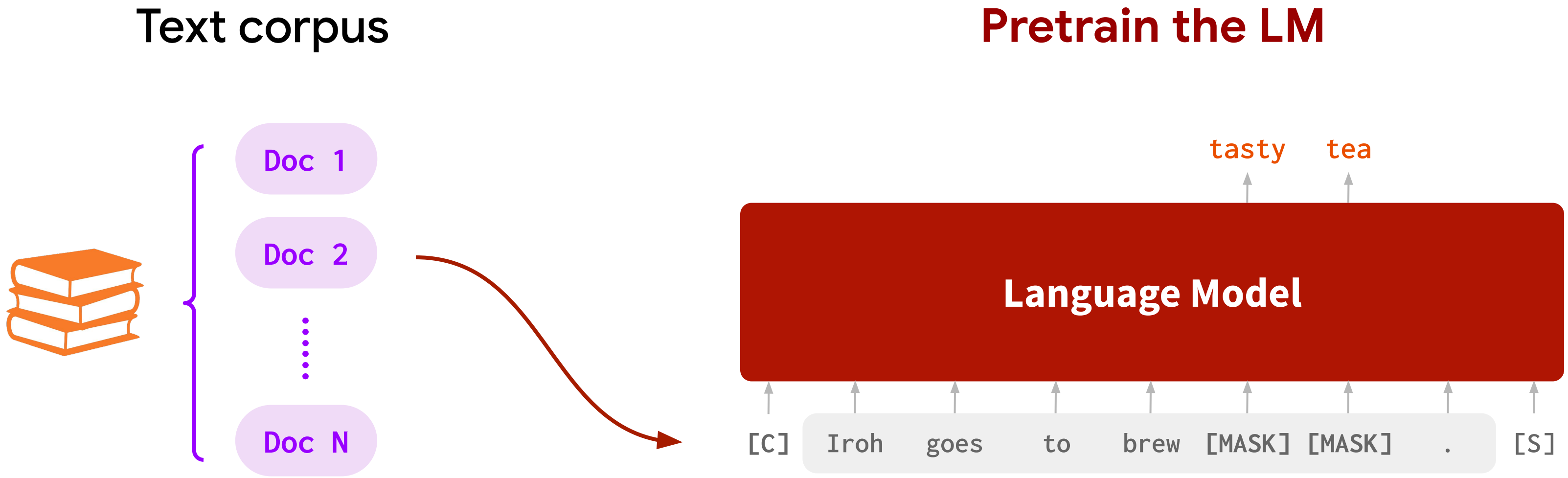
https://www.stateof.ai/

Language Model (LM) Pretraining



Existing LM Pretraining

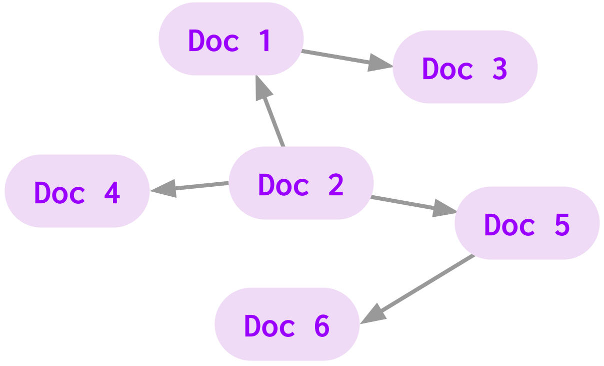
Take a document from text corpus, and perform language modeling over it



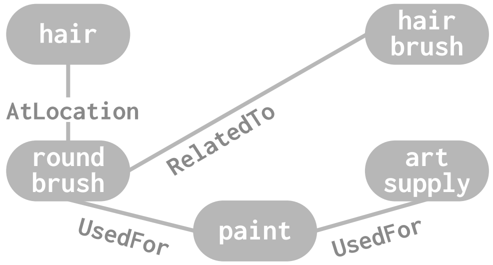
Devlin+2019, Liu+2019

How graphs are useful for LMs?

Hyperlink Graph



Knowledge Graph (KG)



Graphs help make connections between concepts that may be far or latent in text

Graph can bring relevant concepts closer

[Tidal Basin, Washington D.C.]

The Tidal Basin is a man-made reservoir located between .... It is part of West Potomac Park, is near the National Mall and is a focal point of **the National Cherry Blossom Festival** held each spring. The Jefferson Memorial, ...

Document

Hyperlink

[The National Cherry Blossom Festival] ... It is a spring celebration commemorating the March 27, 1912, gift of **Japanese cherry trees** from Mayor of Tokyo City Yukio Ozaki to the city of Washington, D.C. Mayor Ozaki gifted the trees to enhance ...

Linked document

Graph can bring relevant concepts closer

Language model: fine-grained local relations

[Tidal Basin, Washington D.C.]

The Tidal Basin is a man-made reservoir located between .... It is part of West Potomac Park, is near the National Mall and is a focal point of **the National Cherry Blossom Festival** held each spring. The Jefferson Memorial, ...

Document

Graph can provide global relations

Hyperlink

[The National Cherry Blossom Festival] ... It is a spring celebration commemorating the March 27, 1912, gift of **Japanese cherry trees** from Mayor of Tokyo City Yukio Ozaki to the city of Washington, D.C. Mayor Ozaki gifted the trees to enhance ...

Linked document

Multi-hop knowledge

(e.g. *Tidal Basin* has *Japanese cherry trees*)

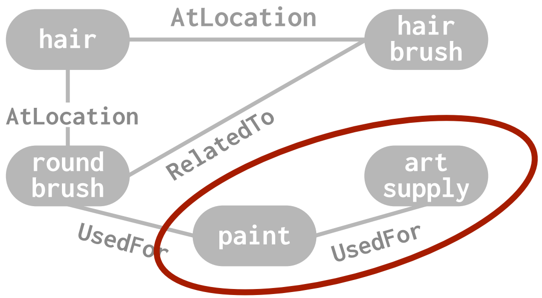
Graph can bring relevant concepts closer

Graph can provide latent relations not mentioned in text

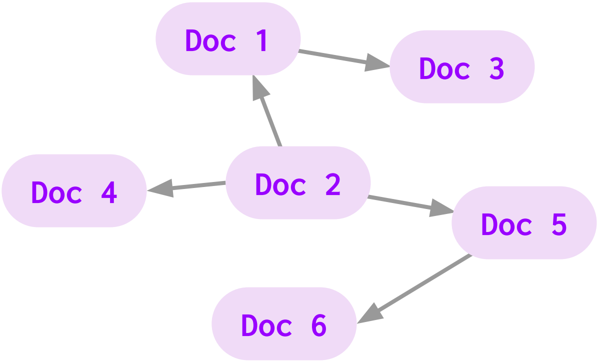
Text

If it is not used for <u>hair</u>,
a <u>round brush</u> can be an
example of what?

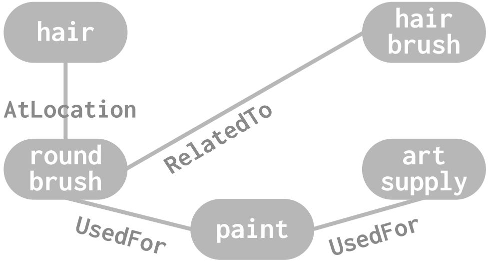
Knowledge Graph



This talk



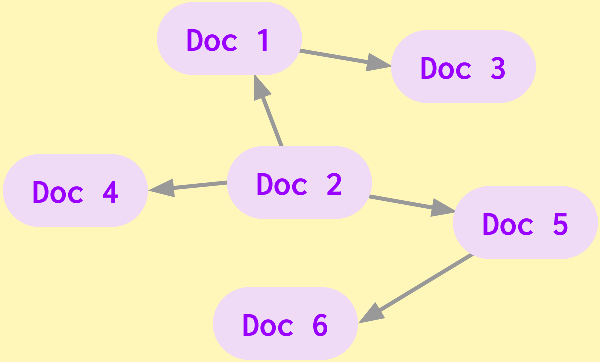
LinkBERT



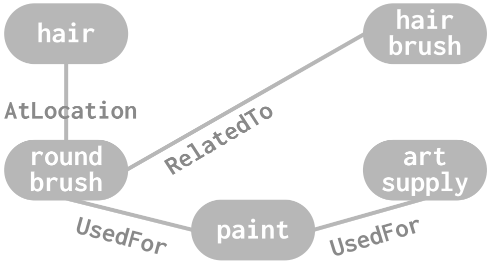
DRAGON

General principle: graphs bring relevant documents/concepts closer together

This talk



LinkBERT



DRAGON

General principle: graphs bring relevant documents/concepts closer together

LinkBERT: Pretraining Language Models with Document Links

ACL 2022

Michihiro Yasunaga, Jure Leskovec*, Percy Liang*
Stanford University

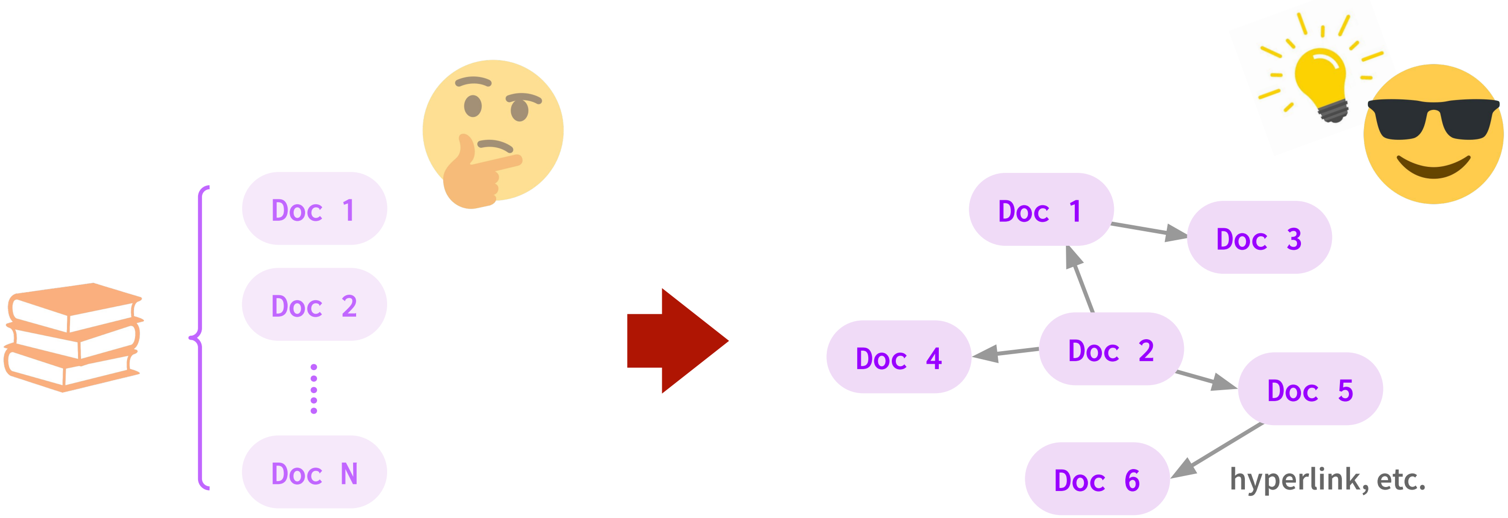
  

But documents have rich dependencies

Corpus is not a list of documents, but a ***graph*** of documents!

Web: **hyperlinks** | Literature: **citations** | Code: **dependencies**

WIKIPEDIA PubMed GitHub



Knowledge can span across documents

[Tidal Basin, Washington D.C.]

The Tidal Basin is a man-made reservoir located between .... It is part of West Potomac Park, is near the National Mall and is a focal point of **the National Cherry Blossom Festival** held each spring. The Jefferson Memorial, ...

Document

Hyperlink

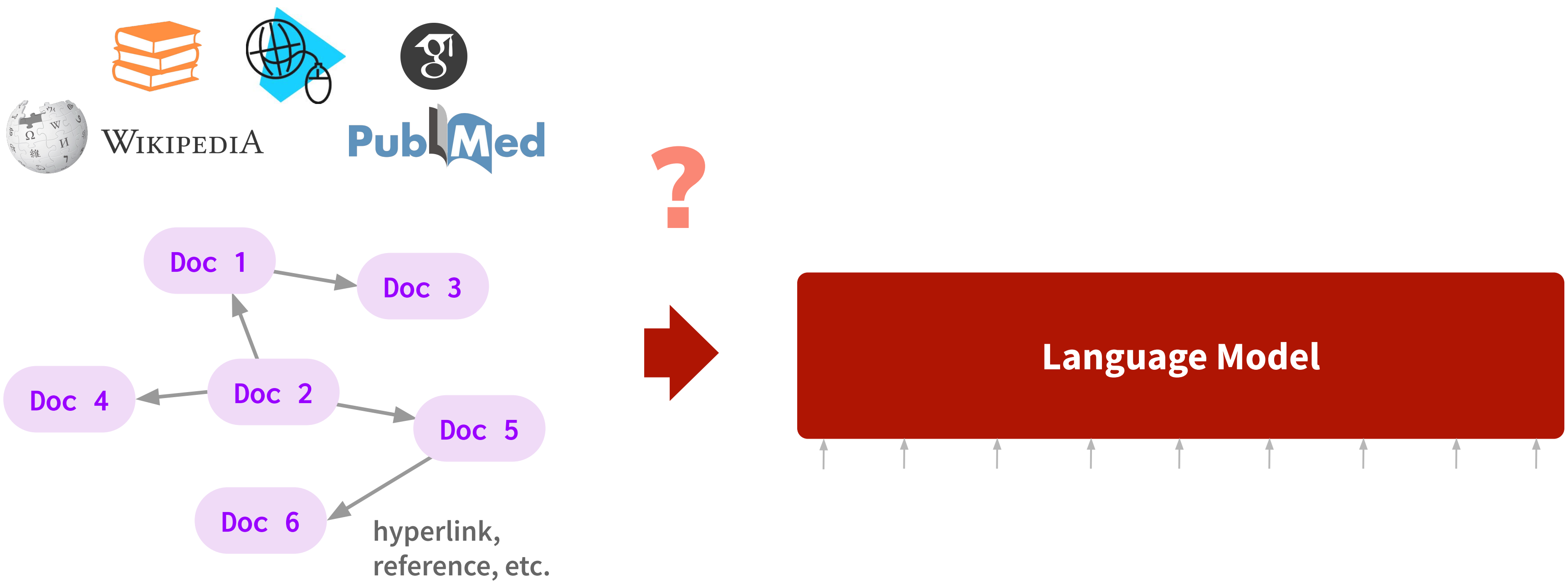
[The National Cherry Blossom Festival] ... It is a spring celebration commemorating the March 27, 1912, gift of **Japanese cherry trees** from Mayor of Tokyo City Yukio Ozaki to the city of Washington, D.C. Mayor Ozaki gifted the trees to enhance ...

Linked document

Multi-hop knowledge

(e.g. *Tidal Basin* has *Japanese cherry trees*)

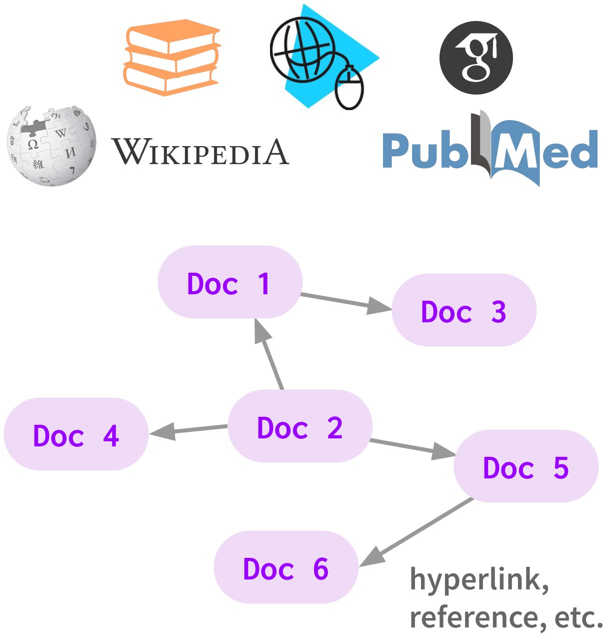
Goal: Train LMs from a Graph of Docs



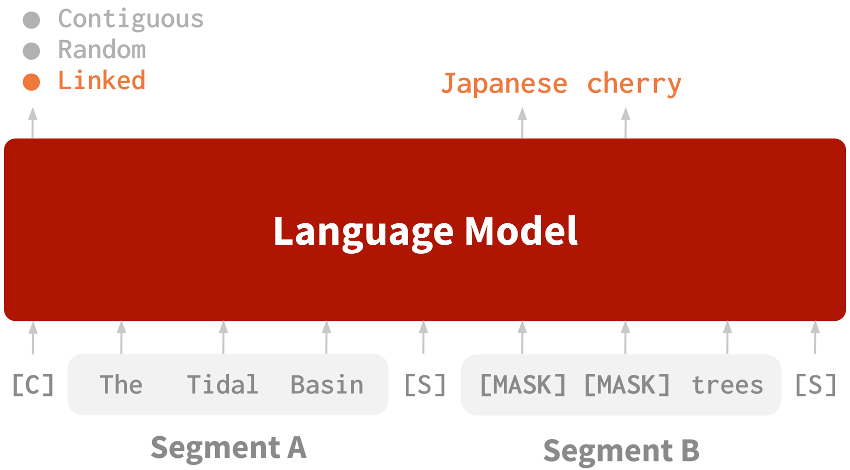
Corpus of linked documents

Pretrain the LM

Proposed Idea: LinkBERT



Corpus of linked documents



Pretrain the LM

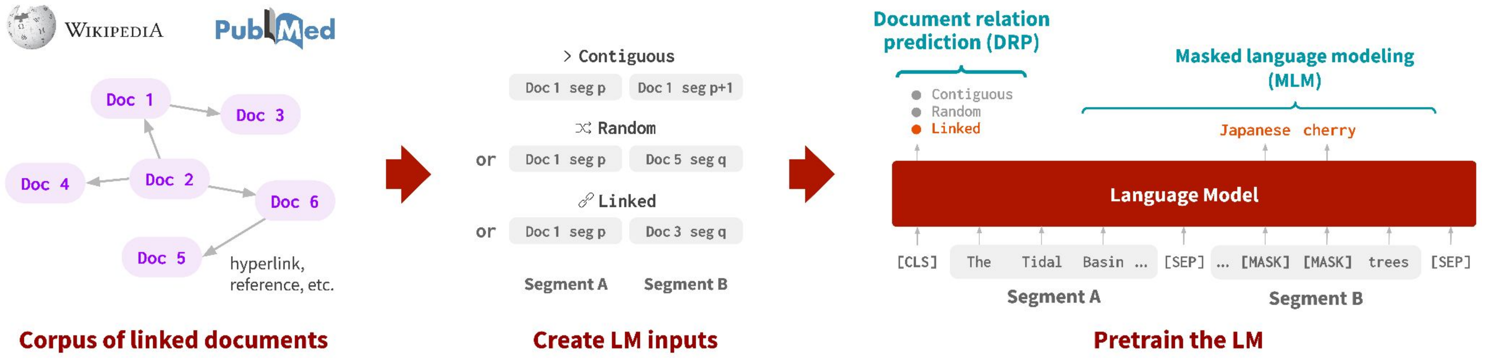
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

(2) Link-aware LM pretraining

- Masked language modeling (MLM)
- Document relation prediction (DRP)



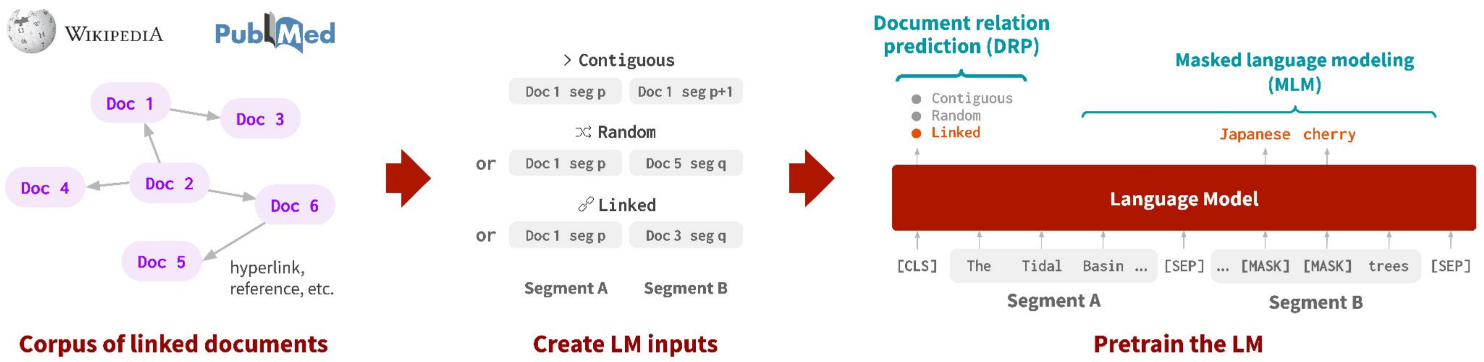
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

(2) Link-aware LM pretraining

- Masked language modeling (MLM)
- Document relation prediction (DRP)



(0) Document Graph

Idea

- Link related docs so that the links can bring together new knowledge

How to link?

- Use **hyperlinks/citations**

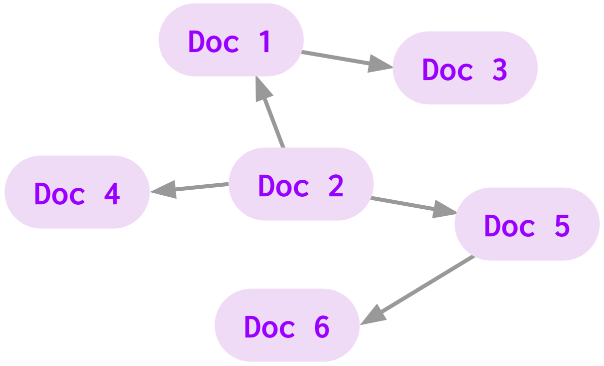
  High quality of relevance. Easily gathered at scale.

- Could also use other linking methods

  e.g. lexical similarity

Build document graph

- Node = document
- Edge (i, j) if there is a link from doc i to doc j



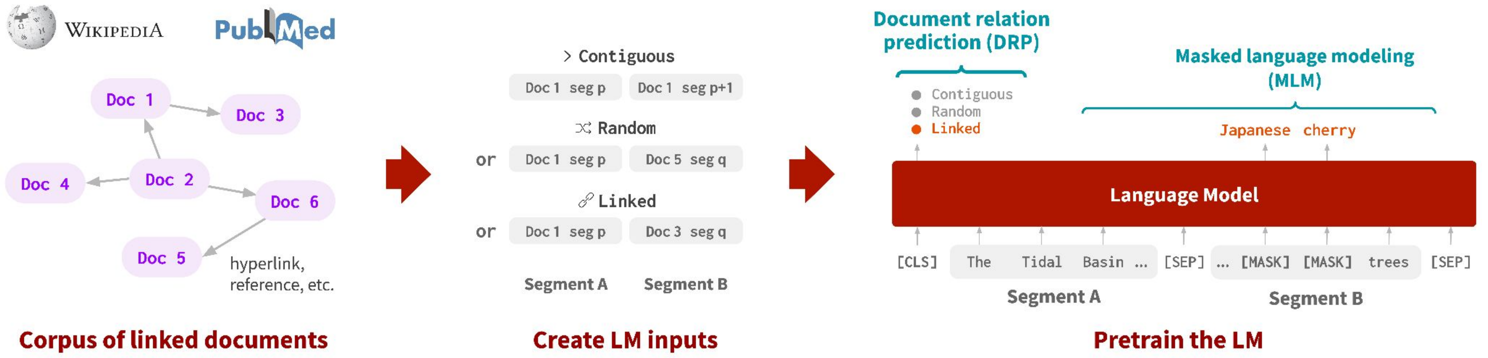
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

(2) Link-aware LM pretraining

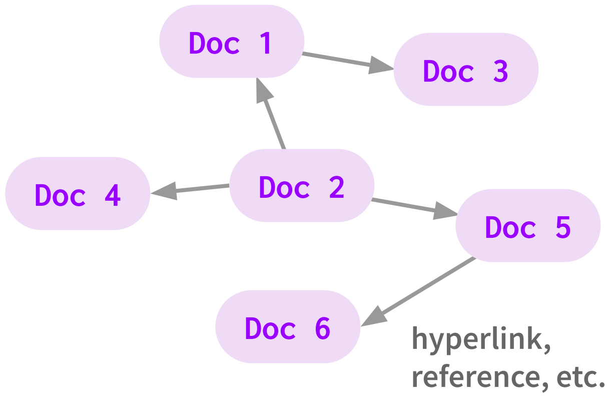
- Masked language modeling (MLM)
- Document relation prediction (DRP)



(1) Link-aware LM Input Creation

Motivation

- LMs learn token dependency effectively if the tokens are shown in the same context (Levine+2022). Let's place linked docs together in the same context 🤝

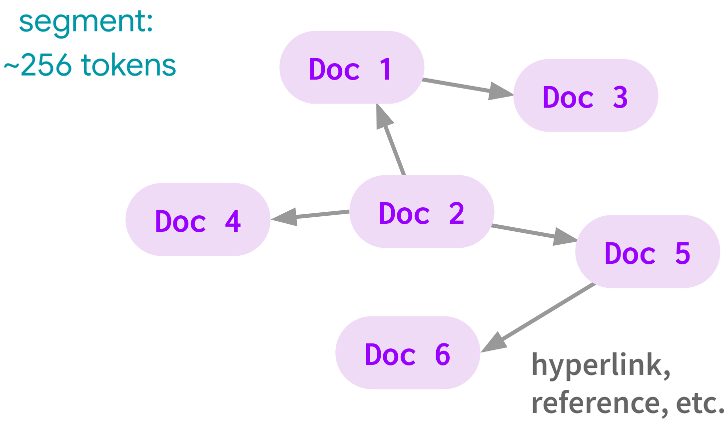


Corpus of linked documents

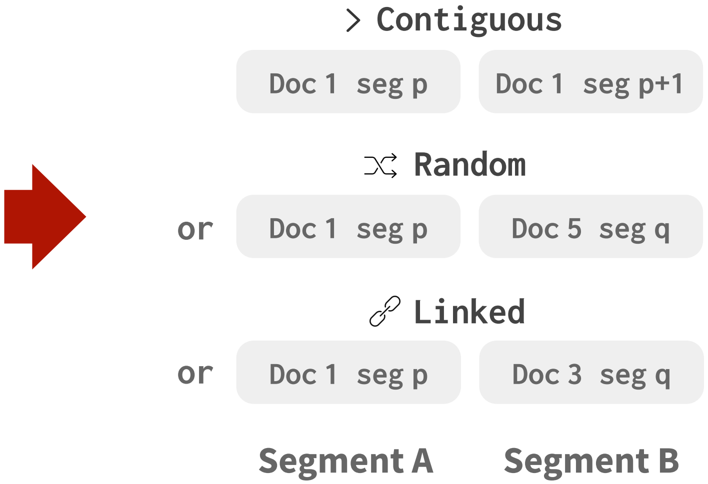
(1) Link-aware LM Input Creation

Idea

- Sample a pair of text segments (A, B) as input, using three options: (i) **contiguous**, (ii) **random**, (iii) **linked**



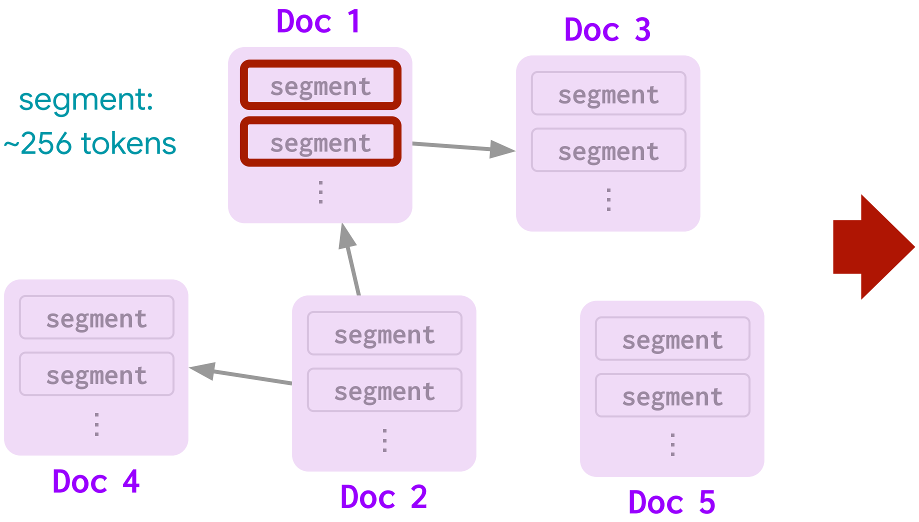
Corpus of linked documents



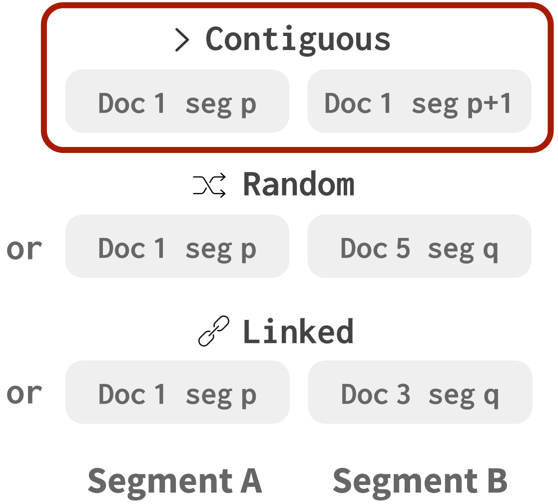
Step 1. Create LM inputs

LM Input Option (i): “Contiguous”

After sampling segment **A**, take the contiguous segment from the same doc as **B** (same as BERT)



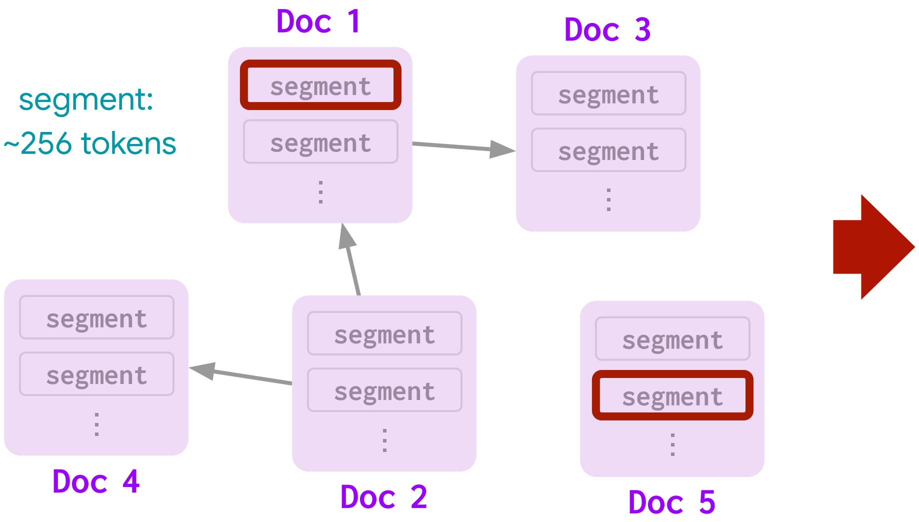
Corpus of linked documents



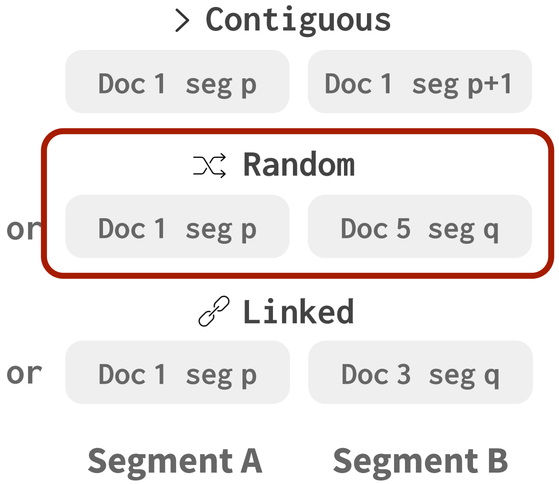
Step 1. Create LM inputs

LM Input Option (ii): "Random"

After sampling segment **A**, sample a segment from a random doc as **B** (same as BERT)



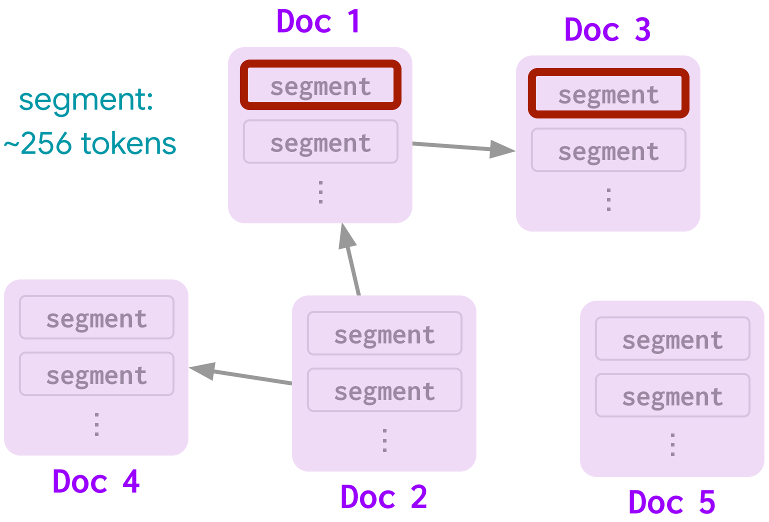
Corpus of linked documents



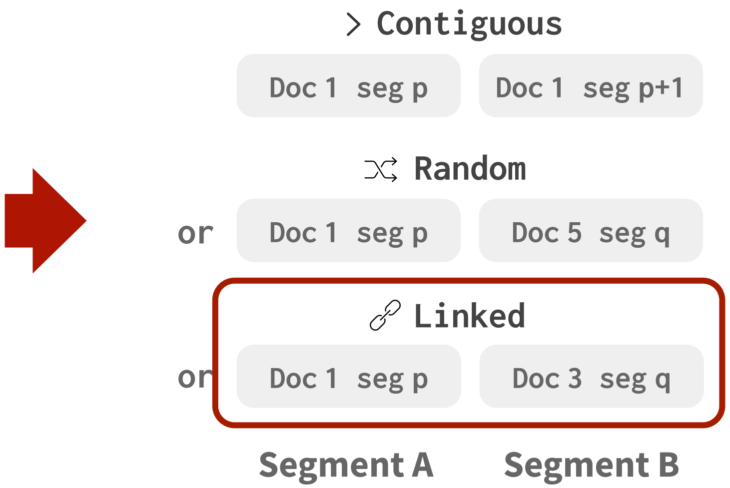
Step 1. Create LM inputs

LM Input Option (iii): “Linked”

After sampling segment **A**, sample a segment from a linked doc as **B**
(our new proposal)



Corpus of linked documents



Step 1. Create LM inputs

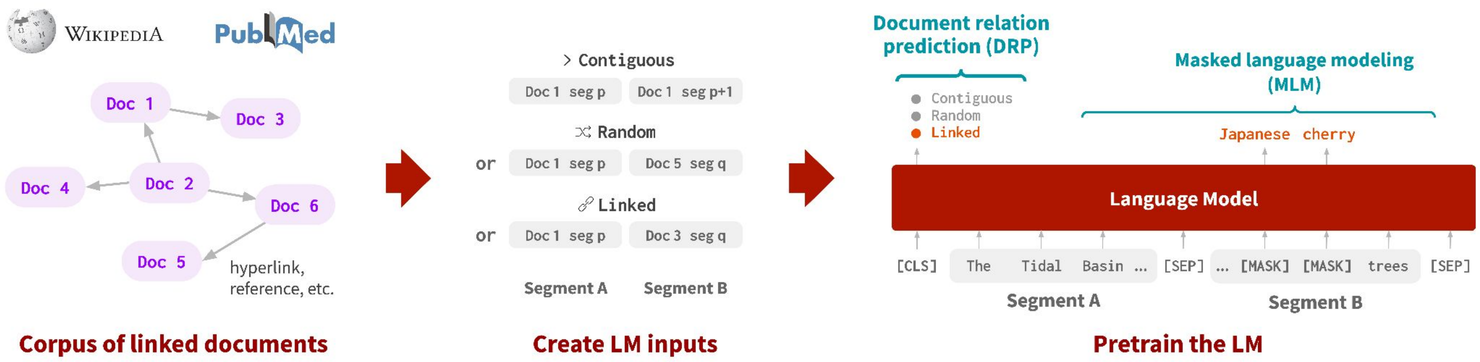
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

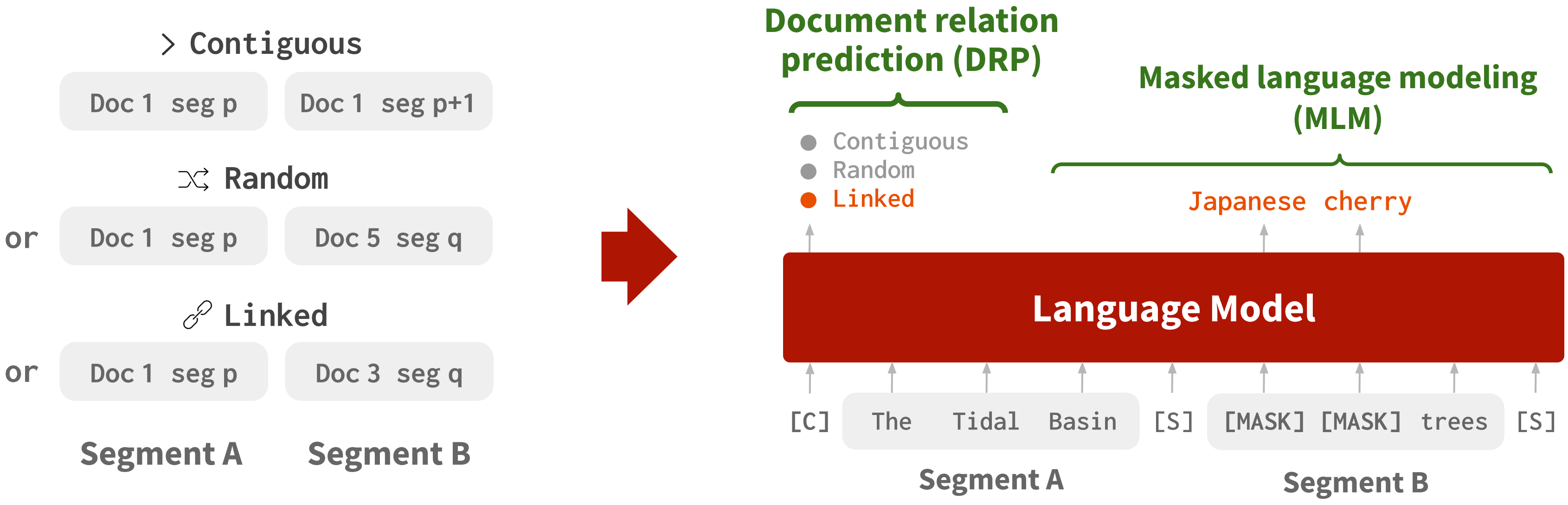
(2) Link-aware LM pretraining

- Masked language modeling (MLM)
- Document relation prediction (DRP)



(2) Link-aware LM Pretraining

Idea: Pretrain LM with link-aware self-supervised tasks



Step 1. Create LM inputs

Step 2. Pretrain the LM

(2) Link-aware LM Pretraining

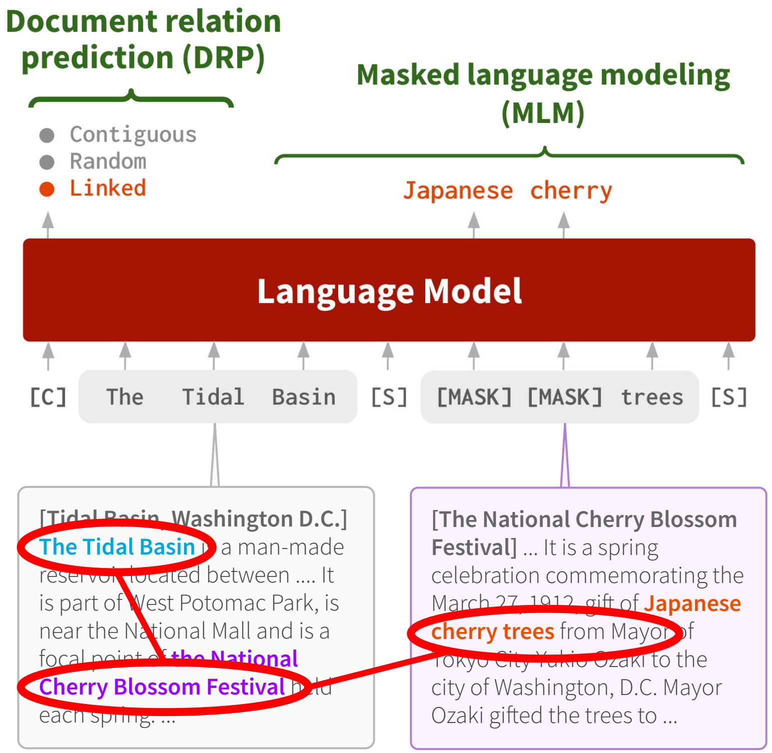
Masked language modeling (MLM)

- Predict masked tokens
- Learn concepts brought into the same context by doc links, e.g. **multi-hop knowlege**

Document relation prediction (DRP)

- Predict the relation between segment A and B
- Learn **relevance** between docs
- Learn the existence of **bridging concepts**

Jointly optimize MLM + DRP



Graph Machine Learning Perspective

Interpretation as graph self-supervised learning on the doc graph

MLM = Node Feature Prediction

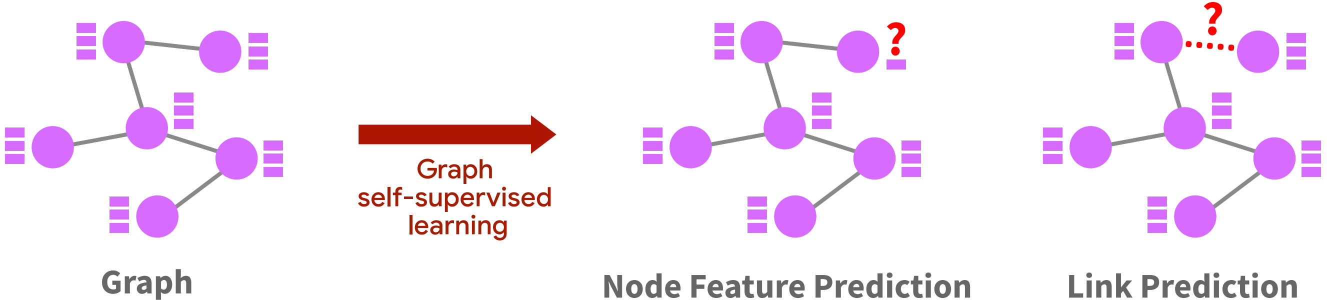
Predict masked features of a node using neighbor nodes

⇒ Predict masked tokens in Segment A using Segment B

DRP = Link Prediction

Predict the existence/type of an edge between two nodes

⇒ Predict if two segments are linked (edge), contiguous (self-loop), or random (no edge)



Bordes+2013, Hu+2020

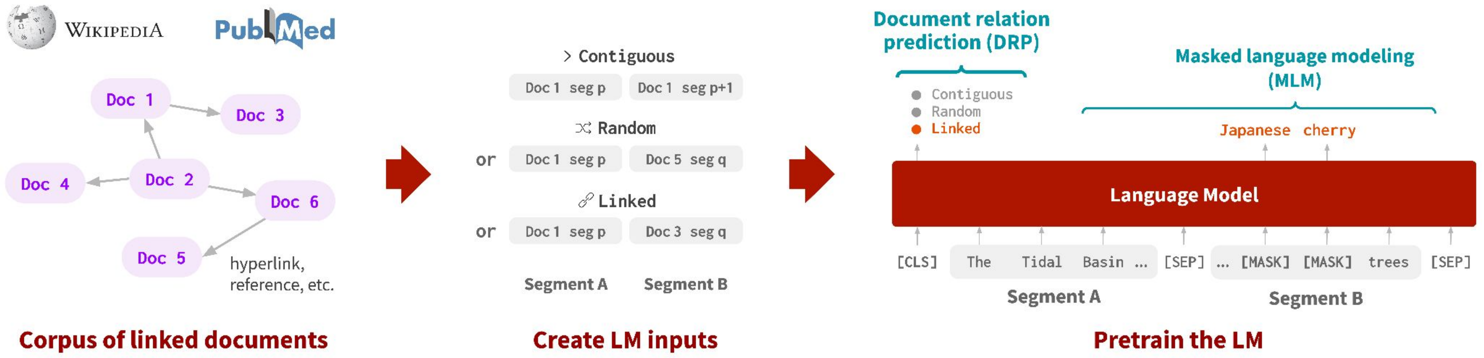
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

(2) Link-aware LM pretraining

- Masked language modeling (MLM)
- Document relation prediction (DRP)

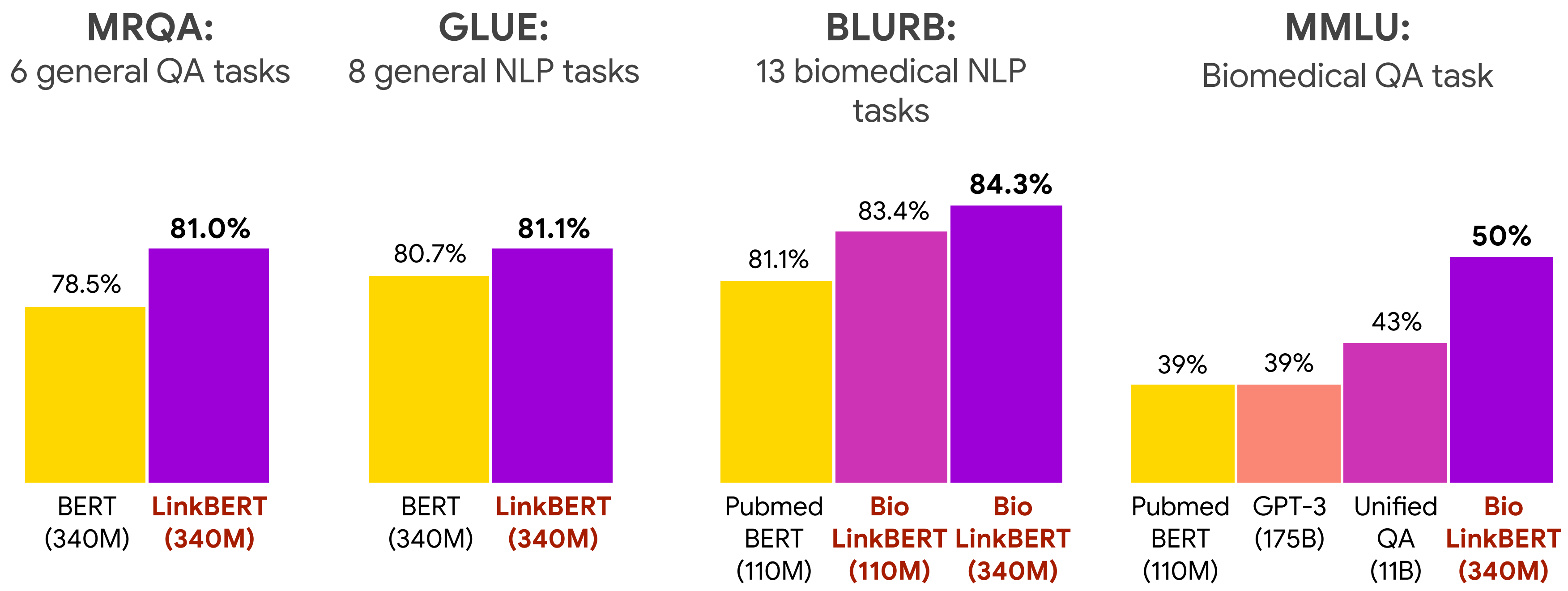


Experiments

	General domain	Biomedical domain
Pretraining corpus	**Wikipedia (10GB) + Books (4GB)** **Links:** hyperlinks **Doc graph:** 3M nodes, 60M edges	**PubMed (20GB)** **Links:** citations **Doc graph:** 15M nodes, 120M edges
Baseline = Pretrained on same corpus, but no doc links	BERT (Devlin+2019)	PubmedBERT (Gu+2020)
Downstream tasks	GLUE (NLP benchmark) MRQA (QA benchmark)	BLURB (NLP benchmark) MedQA-USMLE (QA task) MMLU medicine (QA task)

Performance

LinkBERT makes consistent improvement across tasks and domains



BioLinkBERT sets a new state of the art

BLURB   Leaderboard Paper Models Tasks Submit News

The Overall score is calculated as the macro-average performance over tasks. Details can be found within our publication.

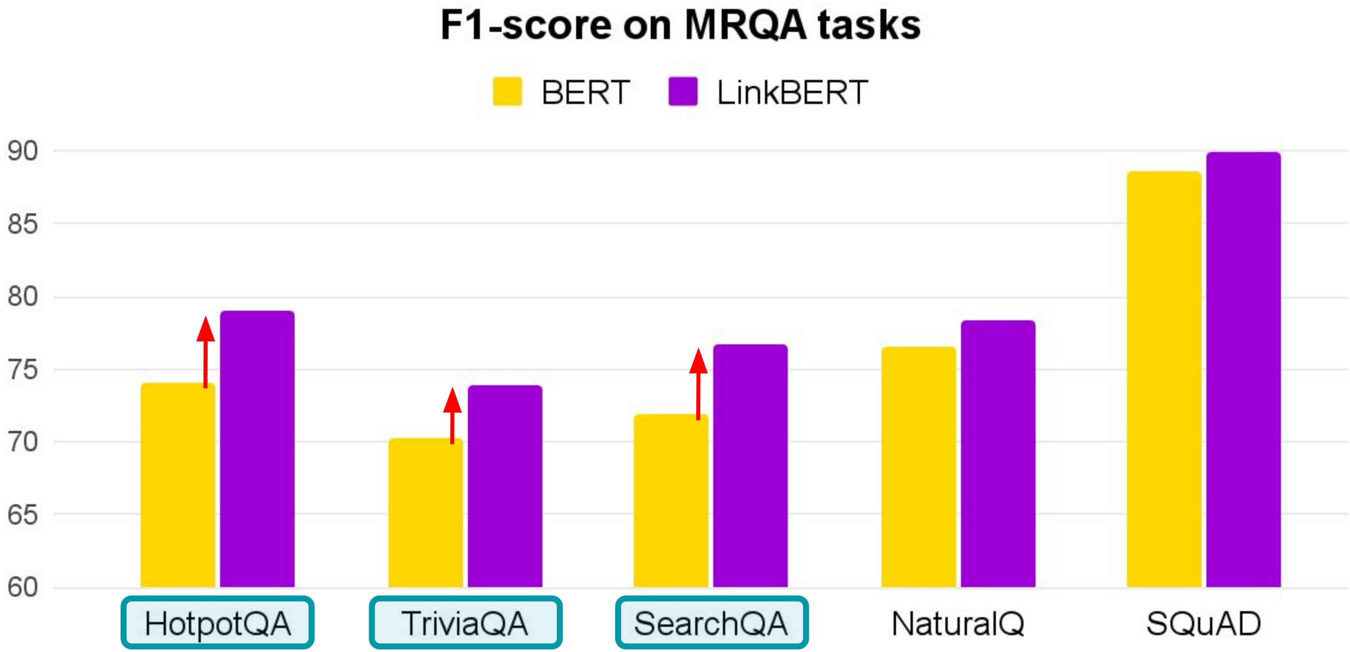
Show 100 entries

Rank	Model	BLURB Score (Macro Avg.)	Micro Avg.	NER	PICO	RE	SS	Class.	QA
1	**BioLinkBERT-Large** — Stanford	**84.30**	**84.80**	**86.89**	**74.19**	**82.74**	93.63	**84.88**	**83.50**
2	**BioLinkBERT-Base** — Stanford	83.39	83.84	86.39	73.97	81.56	93.27	84.35	80.81
3	**PubMedBERT-LARGE (fine-tuning stabilization; uncased; abstracts)** — Microsoft Research	82.91	83.58	86.28	73.61	81.77	92.73	82.70	80.37

https://microsoft.github.io/BLURB/leaderboard.html

Benefit 1: Multi-hop Reasoning

Large gains over BERT on tasks involving multi-hop reasoning



Benefit 1: 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" (✗)**

Intuition: seeing linked docs in the same context in pretraining helps reasoning with multiple docs in downstream

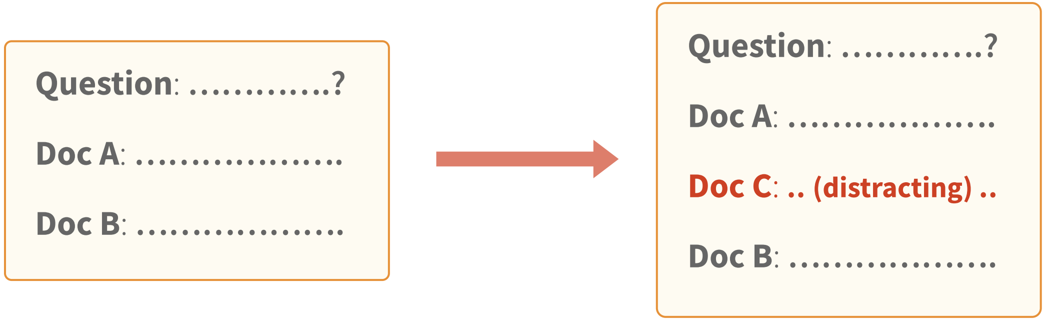
Benefit 2: Document Relation Understanding

Motivation

- In open-domain QA, QA model is given multiple retrieved (**noisy**) documents and needs to understand their relevance (Chen+2017)

Evaluation

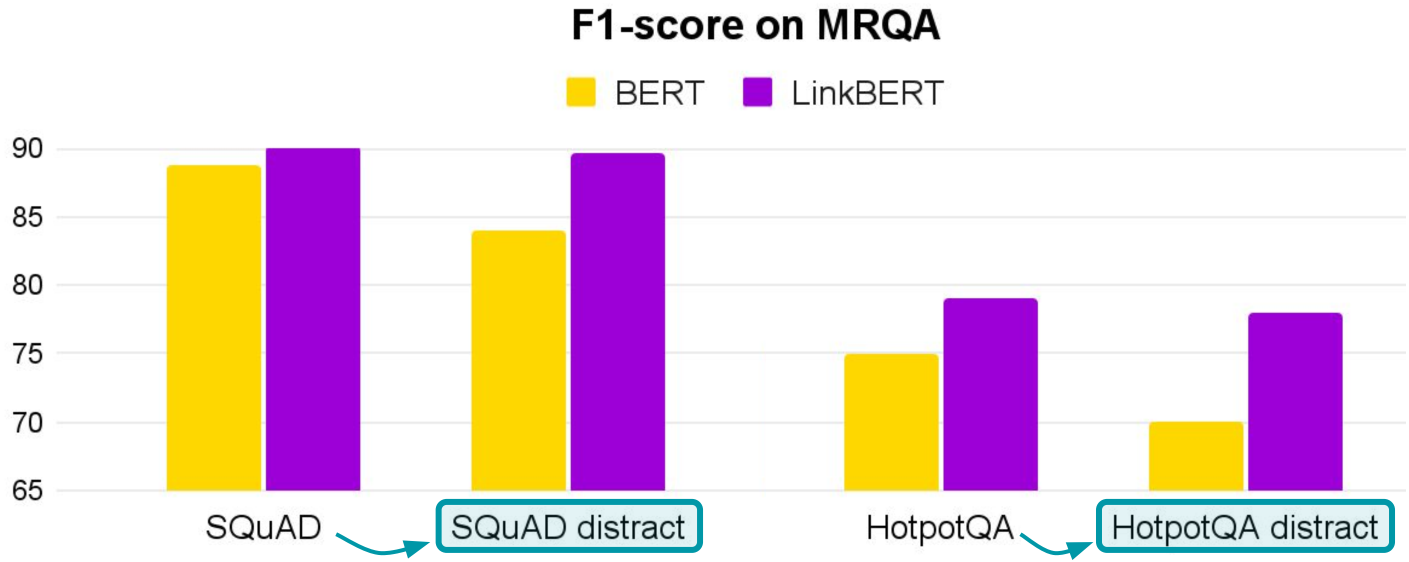
- Add distracting documents to the original MRQA datasets.
  Can LinkBERT still answer correctly?



Benefit 2: Document Relation Understanding

LinkBERT is robust to irrelevant documents

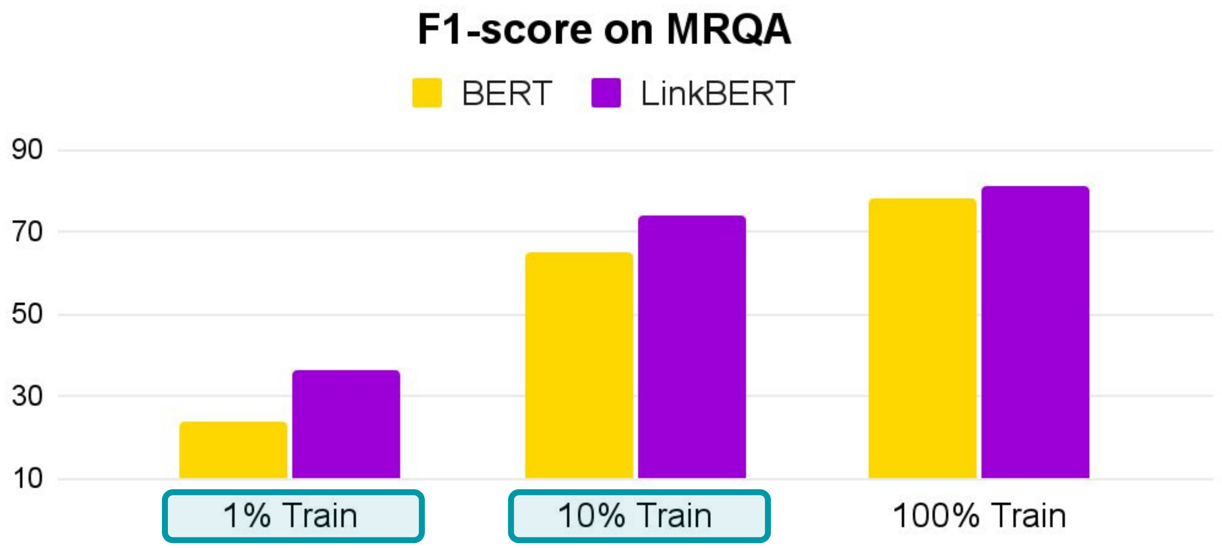
⇒ DRP task in pretraining helps recognizing doc relevance in downstream



Benefit 3: Few-shot QA

Large gains over BERT on few-shot and data-efficient QA

⇒ LinkBERT internalized more knowledge during pretraining



Try our models!

You can easily use LinkBERT on 🤗HuggingFace!

How to use

To use the model to get the features of a given text in PyTorch:

```
from transformers import AutoTokenizer, AutoModel
tokenizer = AutoTokenizer.from_pretrained('michiyasunaga/LinkBERT-large')
model = AutoModel.from_pretrained('michiyasunaga/LinkBERT-large')
inputs = tokenizer("Hello, my dog is cute", return_tensors="pt")
outputs = model(**inputs)
last_hidden_states = outputs.last_hidden_state
```

Takeaways

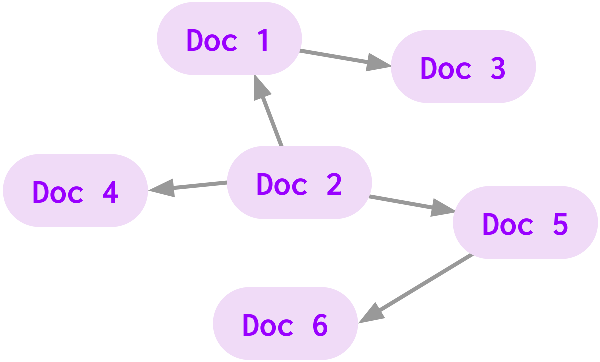
LinkBERT: train LMs using document links (hyperlinks, citations)

Benefits

- Better captures document/concept relations
  ⇒ Effective for **multi-hop** reasoning and **cross-document** understanding

- Internalizes more world knowledge
  ⇒ Effective for **knowledge-intensive** tasks

This talk



LinkBERT

General principle: graphs bring relevant documents/concepts closer together

DRAGON: Deep Bidirectional Language-Knowledge Pretraining

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang,
Chris Manning, Percy Liang*, Jure Leskovec*
Stanford University

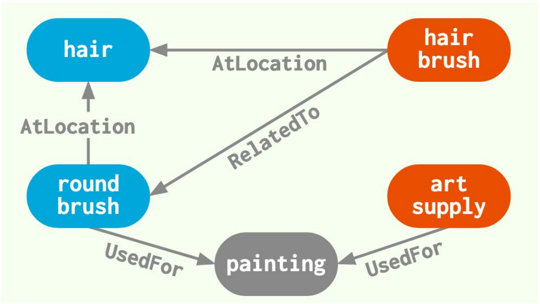
Text & KG offer complementary information

Text & Pretrained Language Model (LM)

- Broad coverage (e.g. Gao+2020)
- Captures rich context

Knowledge Graph (KG)

- Latent, structured relations
- Multihop reasoning (e.g. Yasunaga+2021)



Latent relations about entities that may not be directly mentioned in text

Goal: Combine text & KG for pretraining

Text

- Broad coverage (e.g. Gao+2020)
- Captures rich context

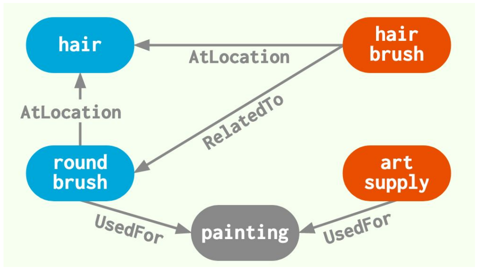
Joint Pretraining

Language-Knowledge Model

Knowledge Graph (KG)

- Latent, structured relations
- Multihop reasoning

(e.g. Yasunaga+2021)

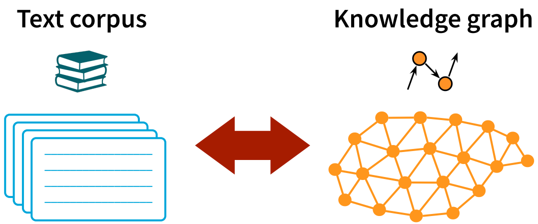


Challenges

How to learn rich representations from text & KG?

(1) Deeply **bidirectional model** for the two modalities to interact

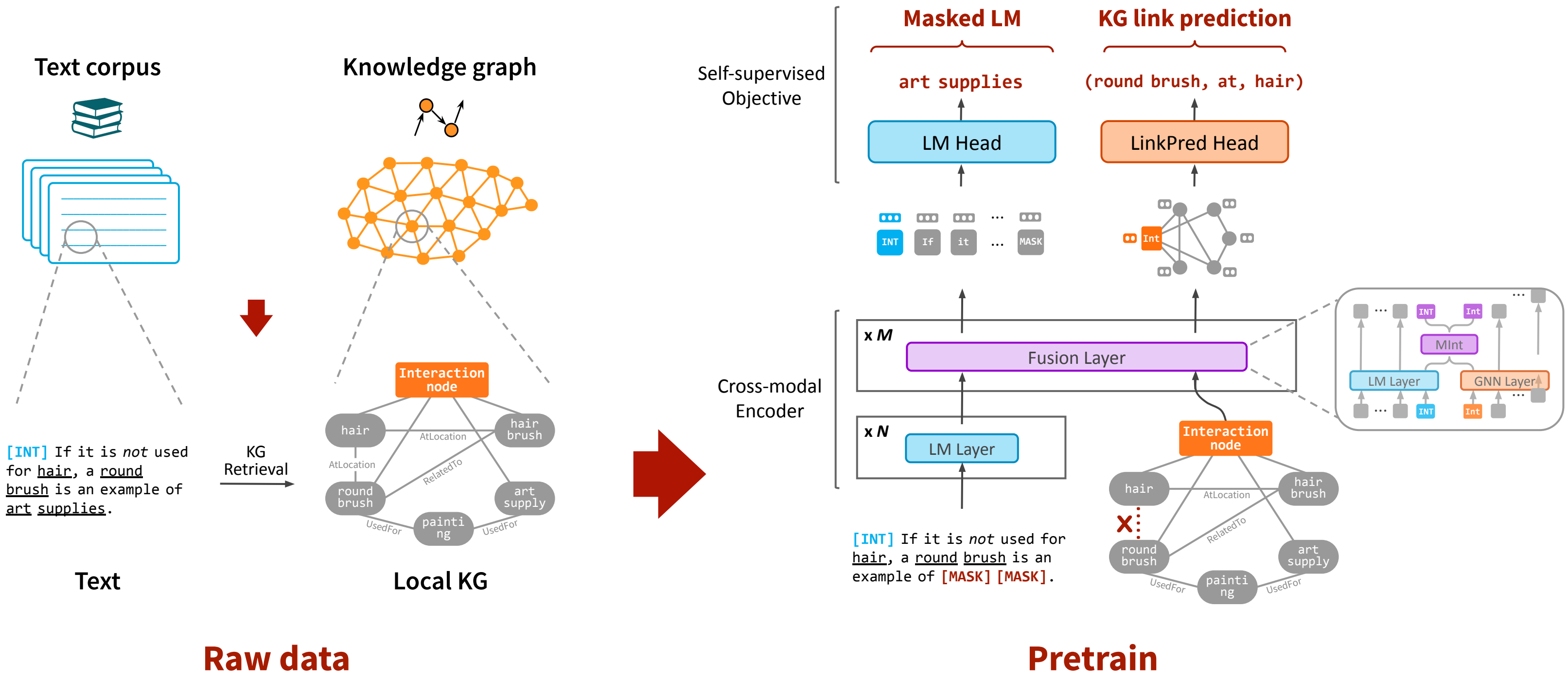
(2) **Self-supervision** to learn joint reasoning over text and KG **at scale**



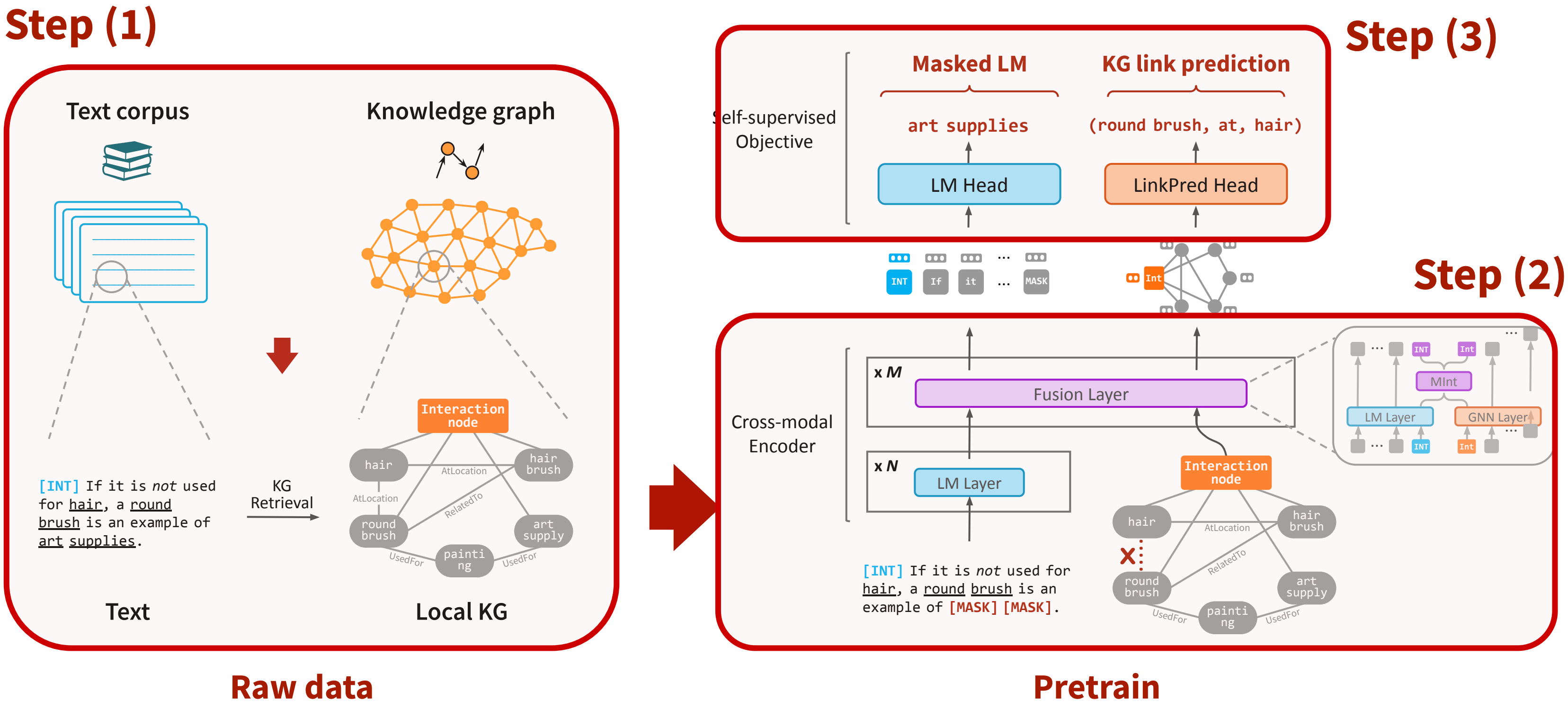
Existing works

- Bidirectional model for text+KG, but only finetune on labeled data (e.g. QAGNN, GreaseLM)
- Self-supervised, but shallow or uni-directional interaction (e.g. ERNIE, WKLM, KEPLER)

Proposed Method: DRAGON



Proposed Method: DRAGON



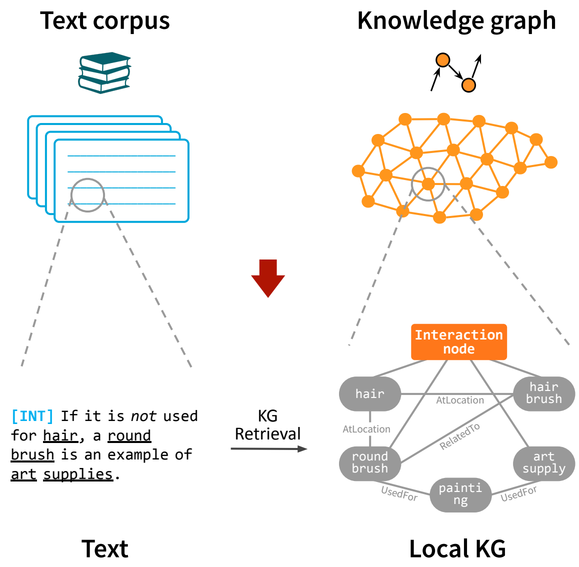
(1) Text-KG Input

Motivation

- Informative (text, local KG) pair:
  - Text can contextualize the KG
  - KG can ground the text

Idea

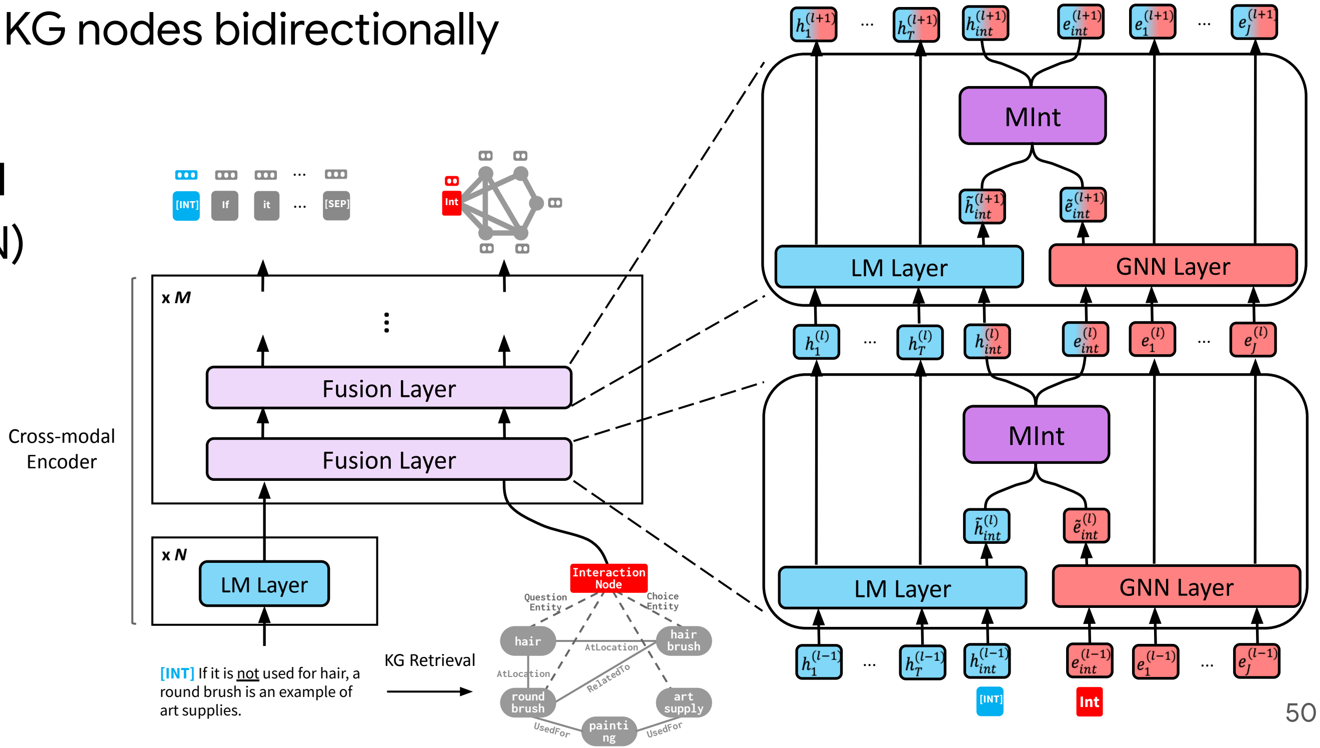
- Given text corpus and KG, create **aligned (text, local KG) pairs** by entity linking and getting neighbors in KG



(2) Deep Bidirectional Cross-Modal Model

Idea

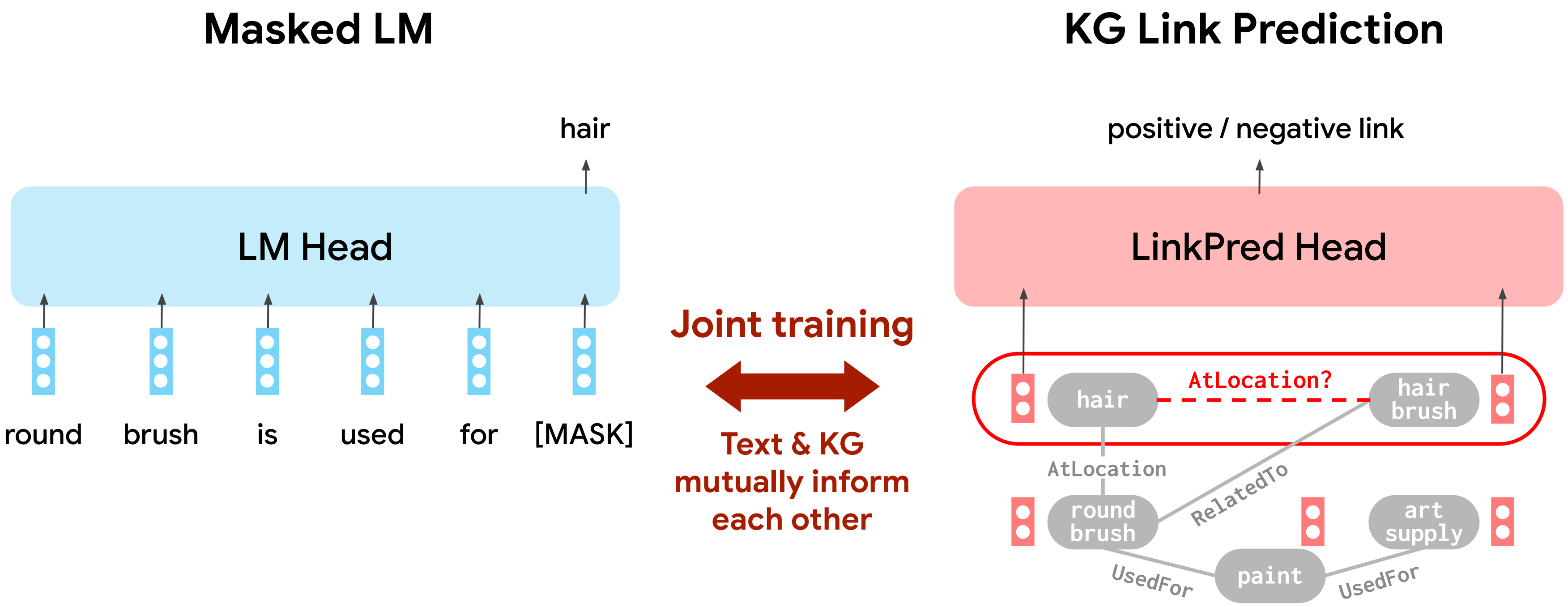
- Fuse text tokens & KG nodes bidirectionally for multiple layers
- Use the **GreaseLM** (Transformer+GNN) encoder



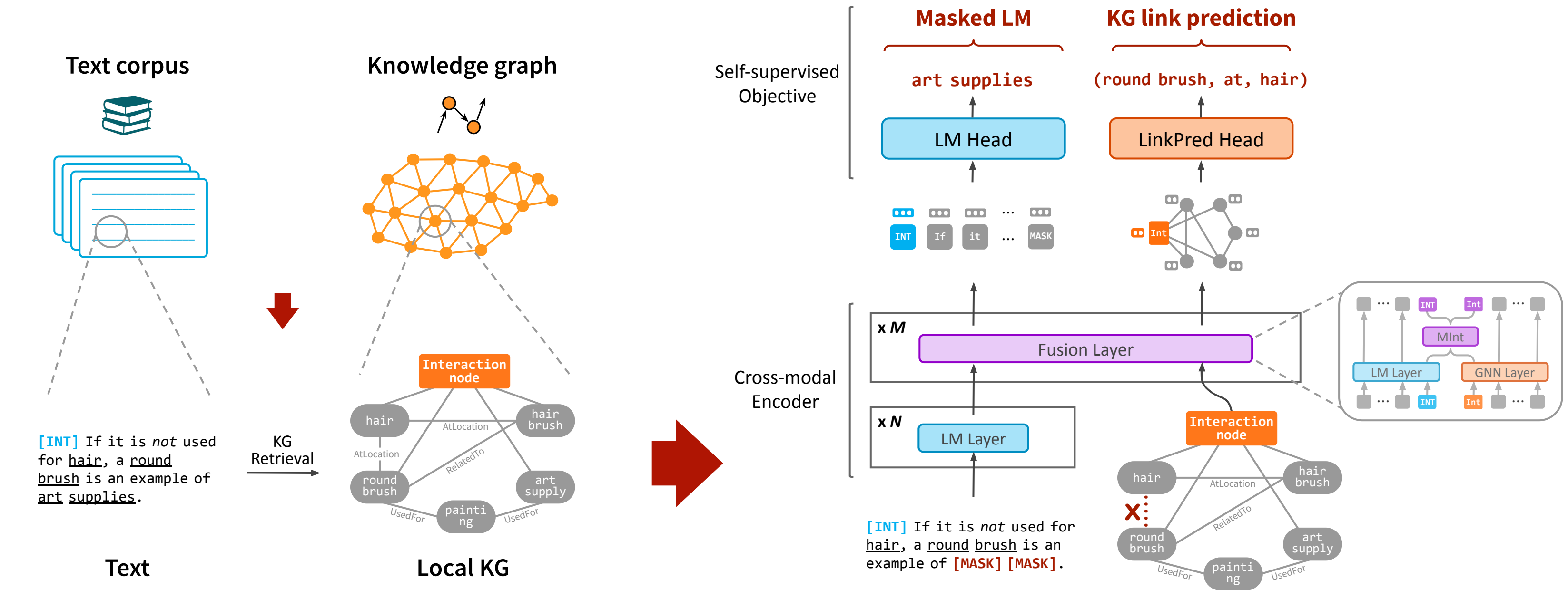
[Zhang et al. 2022](#)

(3) Bidirectional Self-Supervision

Idea: Joint self-supervised objectives



Proposed Method: DRAGON



Raw data

Pretrain

Experiments

	General domain	Biomedical domain
Pretraining data	**Text: BookCorpus** (6GB) **KG: ConceptNet** (800K nodes, 2M edges)	**Text: PubMed** (20GB) **KG: UMLS** (300K nodes, 1M edges)
Downstream tasks	Commonsense reasoning (OBQA, RiddleSense, CommonsenseQA, CosmosQA, HellaSwag, PIQA, SIQA, aNLI, ARC)	Biomedical reasoning (PubMedQA, BioASQ, MedQA-USMLE)
Baseline: LM	RoBERTa (Liu+2019)	BioLinkBERT (Yasunaga+2022)
Baseline: LM finetuned with KG	RoBERTa + GreaseLM	BioLinkBERT + GreaseLM

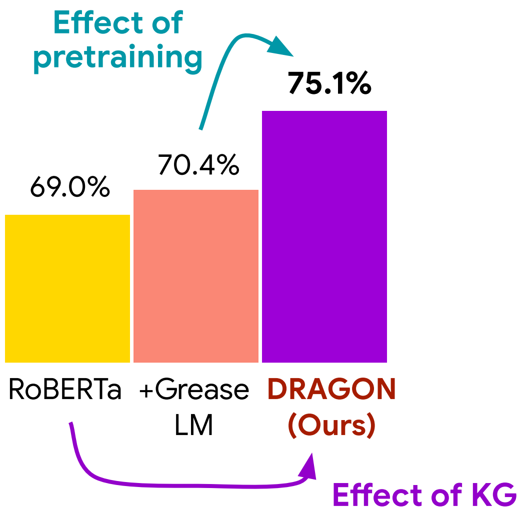
Ours (DRAGON): LM *pretrained* with KG

Performance

DRAGON makes consistent improvement across tasks and domains

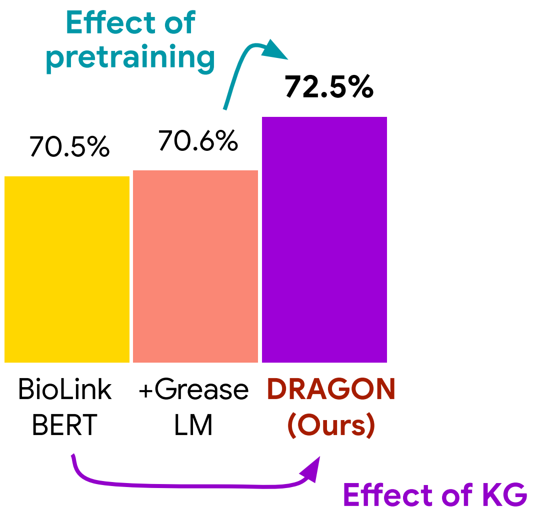
Commonsense reasoning tasks

(e.g. OBQA, RiddleSense)



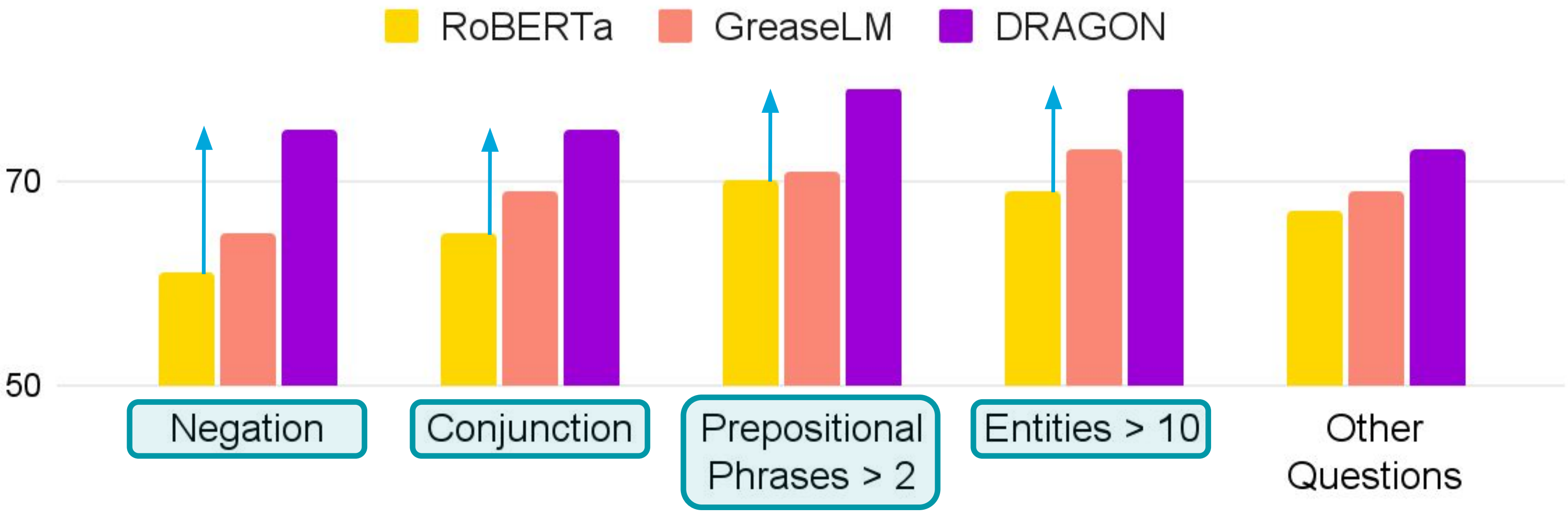
Biomedical reasoning tasks

(e.g. PubMedQA, MedQA)



Benefit: Complex Reasoning

Large gains on QA examples involving complex reasoning

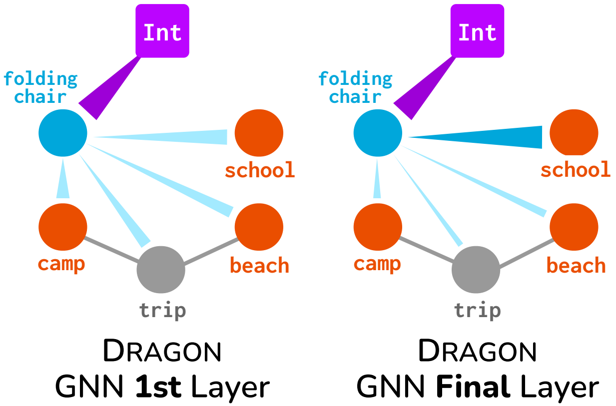


Benefit: Complex Reasoning

Conjunction

Where would you use a **folding chair** **and** store one?

A. camp **B. school** C. beach



RoBERTa:
A. camp (✗)

GreaseLM:
C. camp (✗)

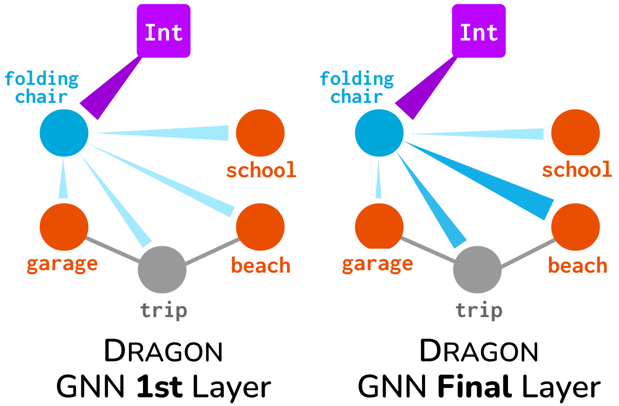
DRAGON:
B. school (✓)

Model Prediction

Negation + Conjunction

Where would you use a **folding chair** **but not** store one?

A. garage B. school **C. beach**



RoBERTa:
B. school (✗)

GreaseLM:
B. school (✗)

DRAGON:
C. beach (✓)

Model Prediction

In DRAGON, KG serves as scaffold for performing structured/multi-step reasoning

Summary

DRAGON: Pretrain a foundation model jointly on text & KGs

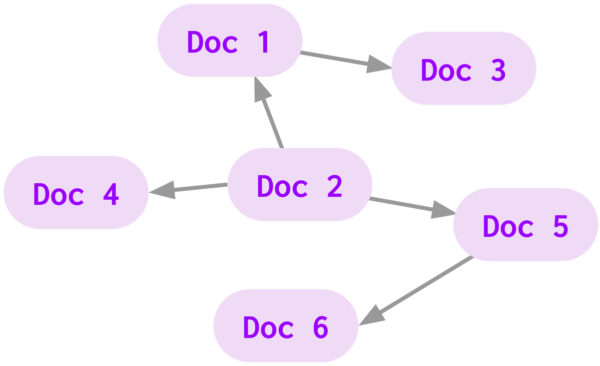
Approach

- Deeply bidirectional model for the two modalities to interact
- Self-supervised objective to learn joint reasoning over text and KG at scale

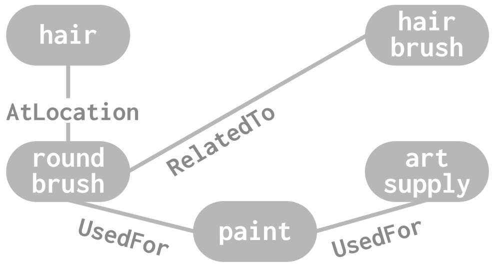
Result

- Improved performance on knowledge- and reasoning-intensive applications (e.g. low-resource QA, multi-step reasoning)

Final remarks



LinkBERT



DRAGON

General principle: graphs bring relevant documents/concepts closer together

Open question: how to better incorporate implicit relations (e.g., entity mentions w/o hyperlinks)

…The campus occupies 8,180 acres (3,310 hectares), among the largest in the United States…

Open question: how to perform more formal reasoning at scale?

References

- Michihiro Yasunaga, Jure Leskovec, Percy Liang.
  LinkBERT: Pretraining Language Models with Document Links. ACL 2022.
- Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, Jure Leskovec.
  QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. NAACL 2021.
- Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Chris Manning, Jure Leskovec.
  GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. ICLR 2022.
- Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Chris Manning, Percy Liang, Jure Leskovec.
  DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining. NeurIPS 2022.

- **Code/Models**
  - https://github.com/michiyasunaga/LinkBERT
  - https://github.com/michiyasunaga/QAGNN
  - https://github.com/michiyasunaga/dragon

Collaborators