LinkBERT: Pretraining Language Models with Document Links

ACL 2022

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

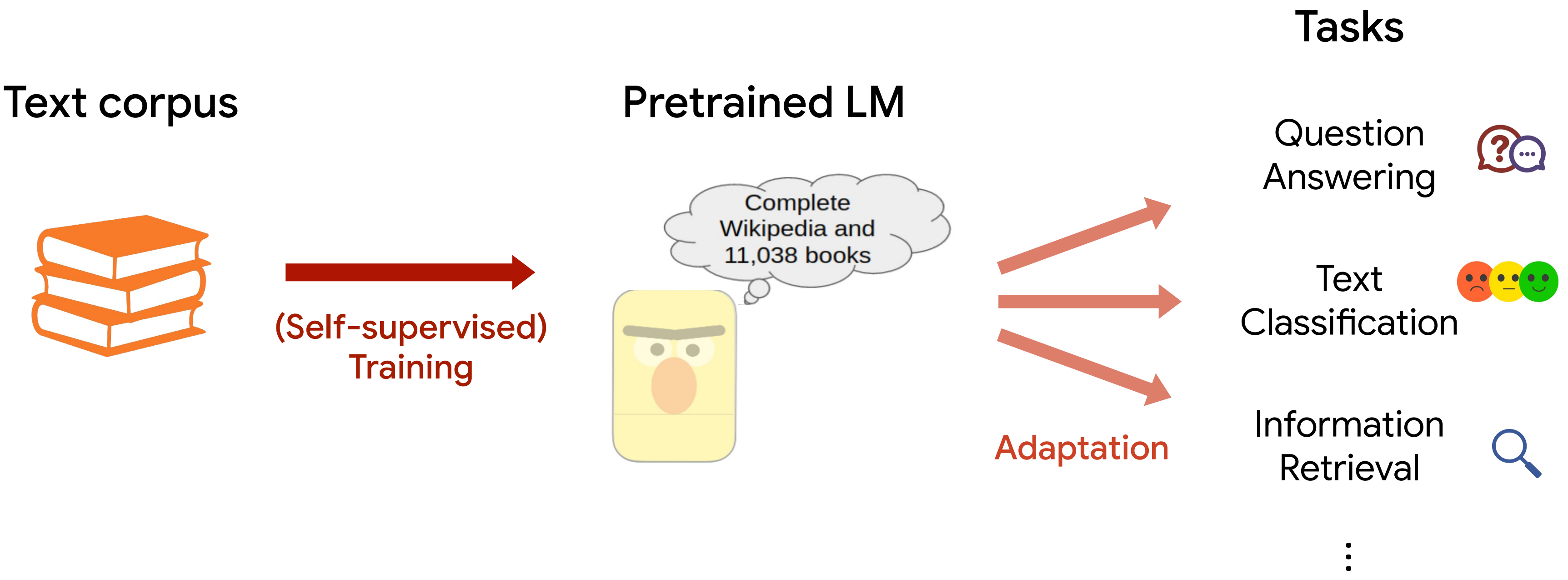
Center for Research on Foundation Models

Stanford NLP

Stanford Artificial Intelligence

Language Model (LM) Pretraining

Core component of today's NLP systems



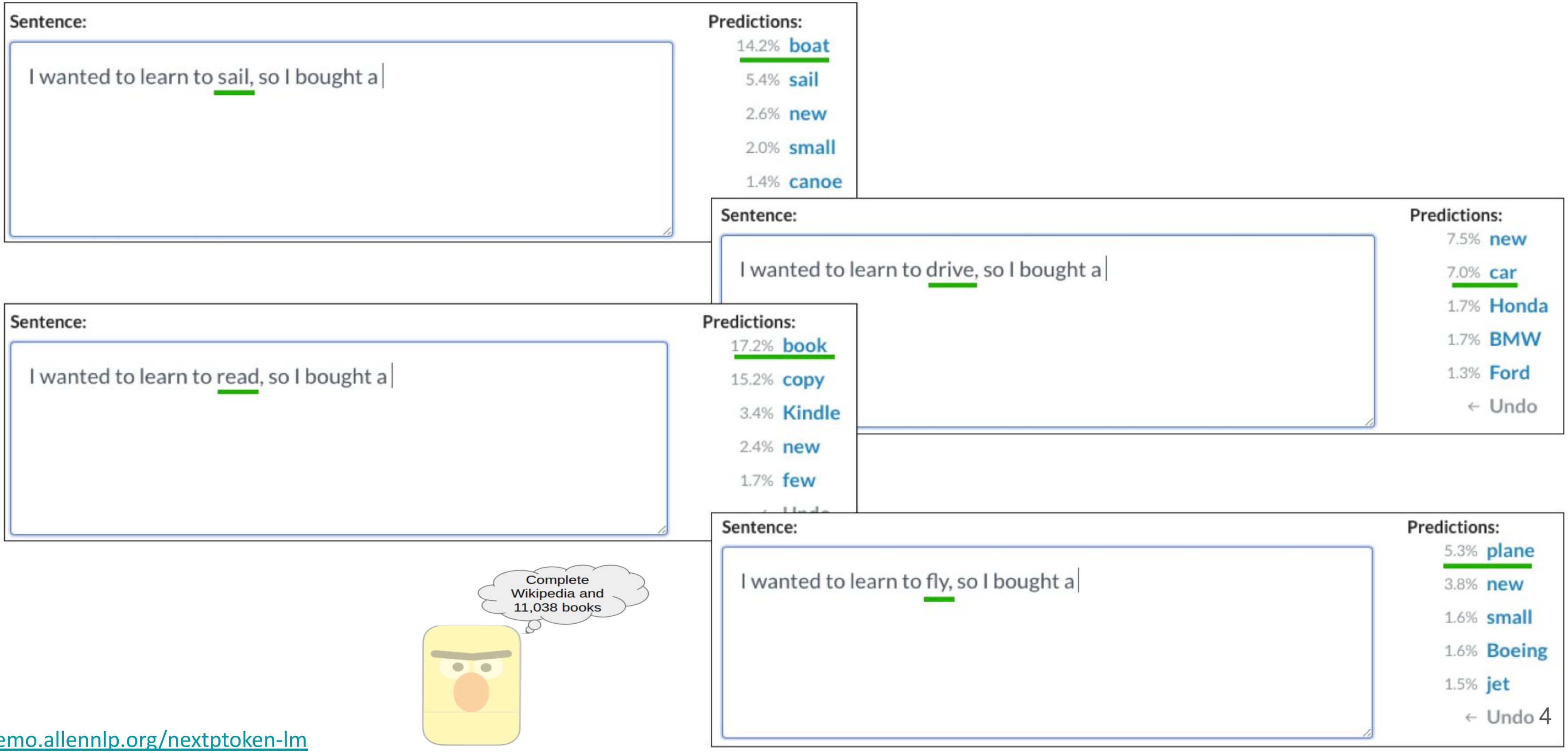
Bommasani+2021

Language Model (LM) Pretraining

Large-scale self-supervised learning

Task	Examples	Input	Output
Masked LM	BERT, RoBERTa, etc.	`‘My dog is fetching the’`	`next_word = ‘ball’`
Causal LM	GPT-*	`‘My __ is fetching the ball’`	`mask = ‘dog’`
Seq2seq	BART, T5, etc.	`‘My __ is fetching the ball’`	`denoised = ‘My dog is fetching the ball’`

LMs learn various knowledge



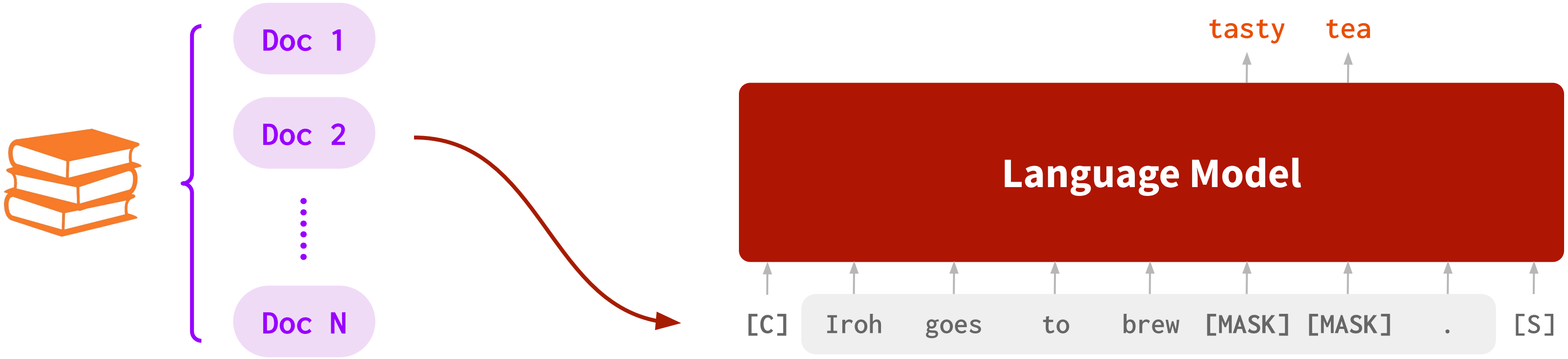
https://demo.allennlp.org/nextptoken-lm

Existing LM Pretraining Methods

Tyically model a **single** document at a time (e.g. BERT, RoBERTa)

Text corpus

Pretrain the LM



Devlin+2019, Liu+2019

But documents have rich dependencies

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

Web: **hyperlinks**

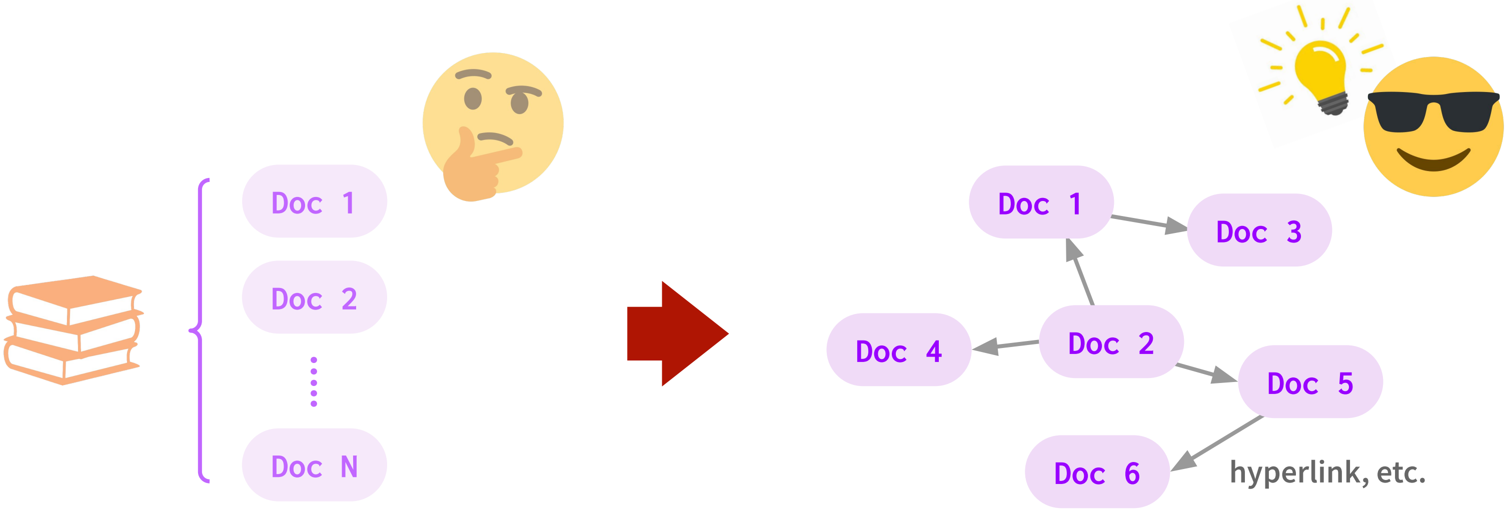
WIKIPEDIA

Literature: **citations**

PubMed

Code: **dependencies**

GitHub



Knowledge can span across documents

Document - - - - - - - ► **Linked document**

(e.g. hyperlink, citation)

[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, ...

[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 ...

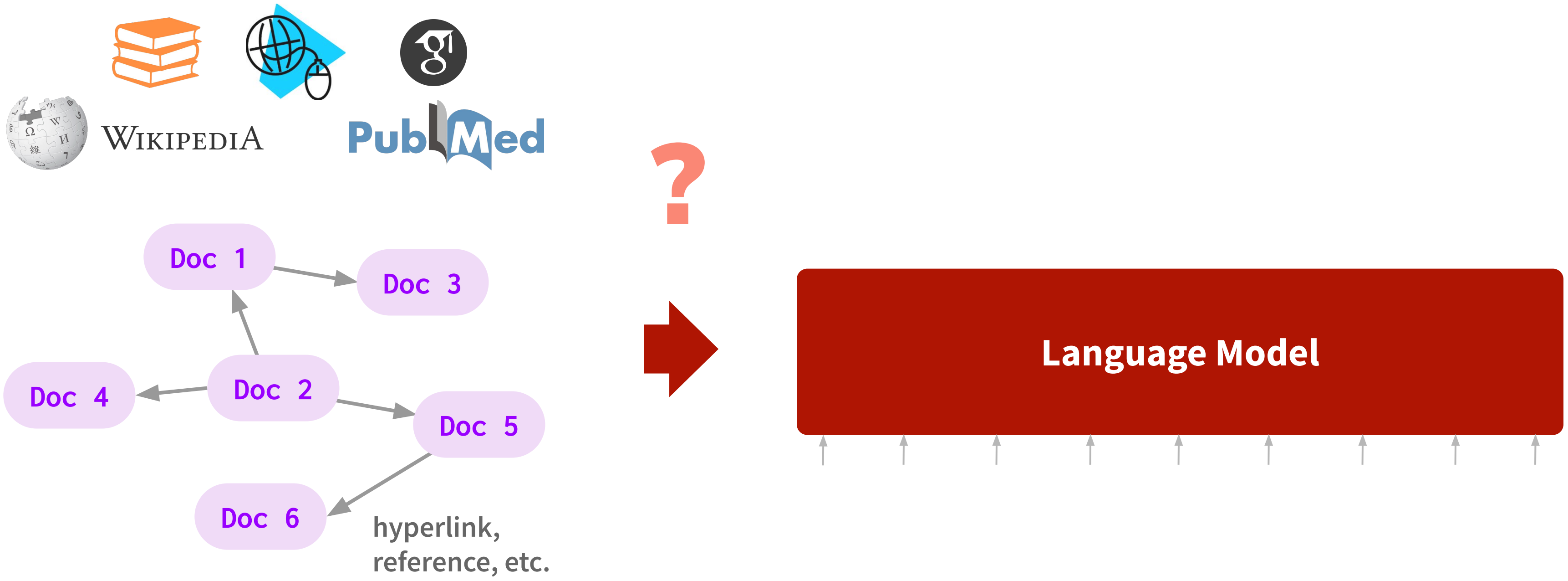
Multi-hop knowledge

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

Document links offer **new knowledge** not available in single documents alone.

Useful for **various applications**, e.g. QA, discovery.

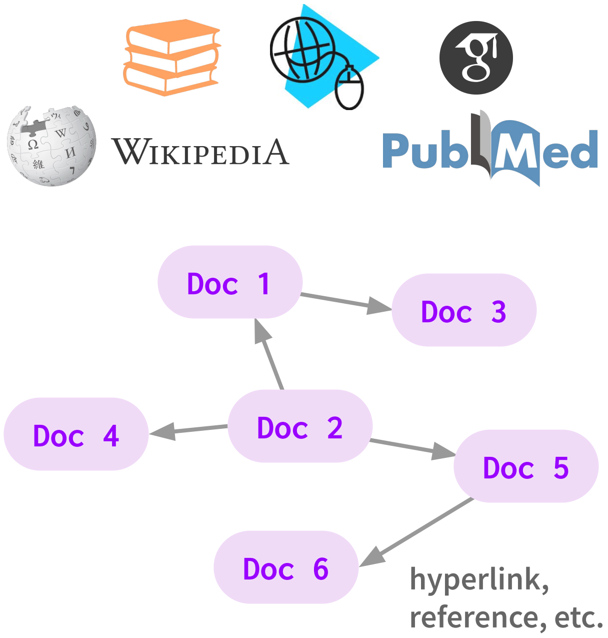
Goal: Train LMs from a Graph of Docs



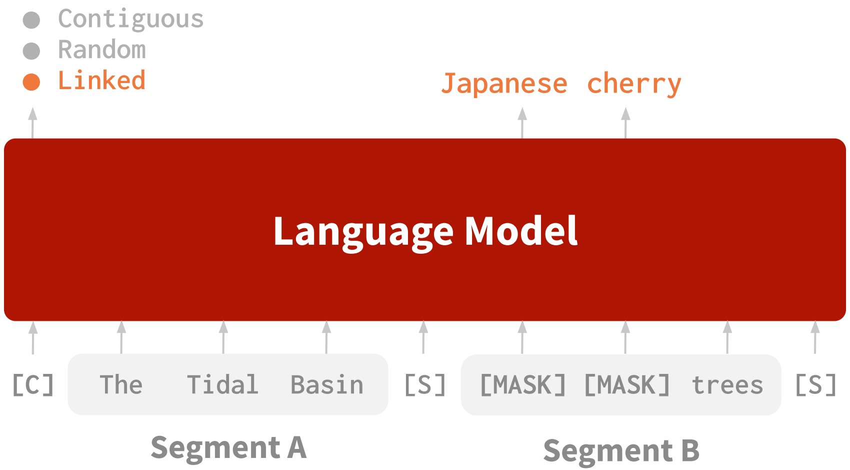
Corpus of linked documents

Pretrain the LM

Proposed Idea: LinkBERT



Corpus of linked documents



Pretrain the LM

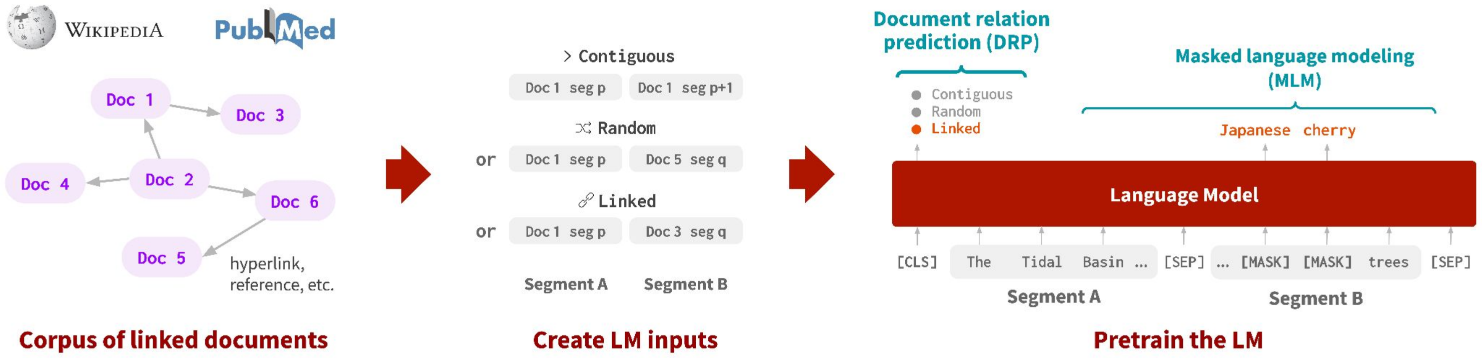
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

(2) Link-aware LM pretraining

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



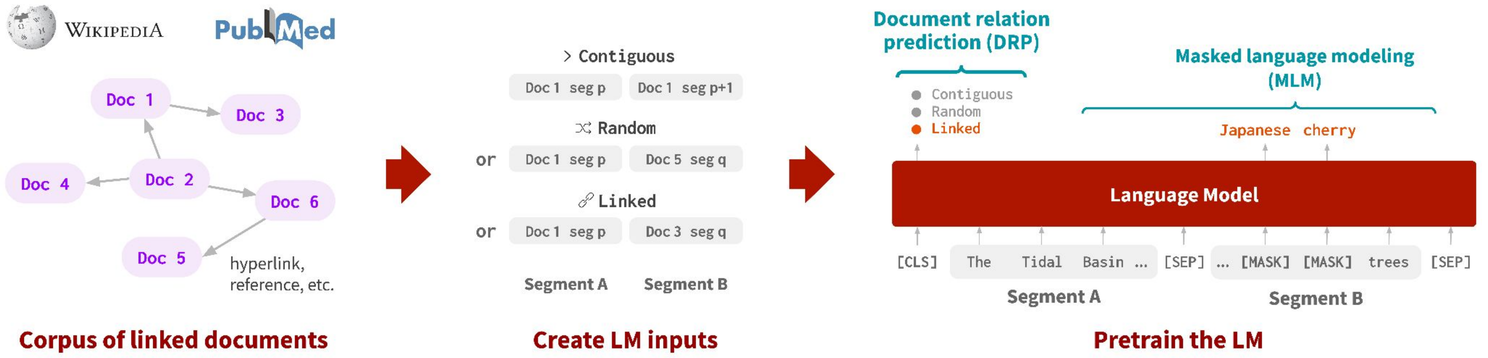
Proposed Idea: LinkBERT

(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

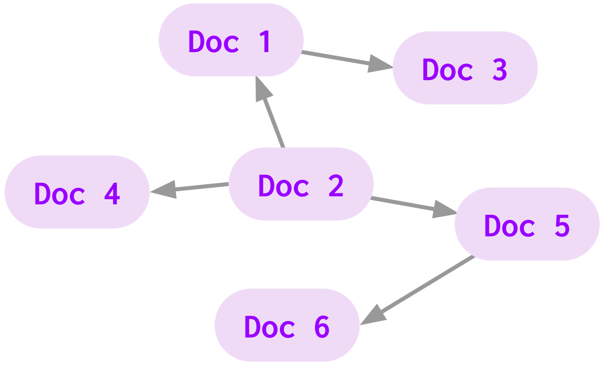
How to link?

- 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

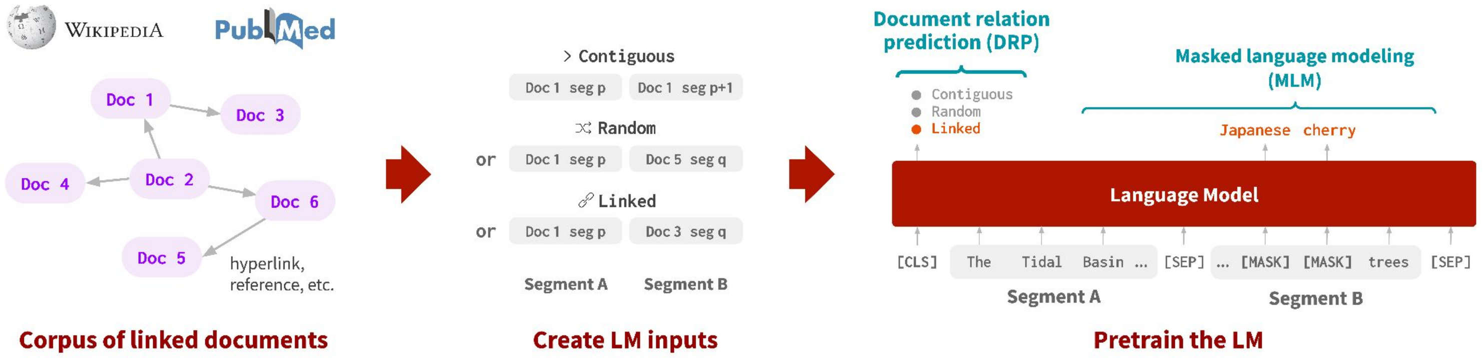
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

(2) Link-aware LM pretraining

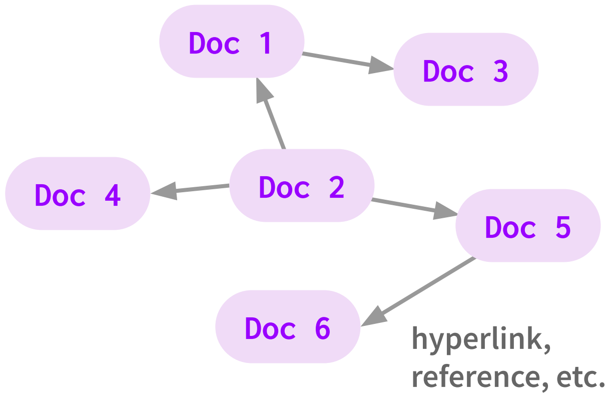
- Masked language modeling (MLM)
- Document relation prediction (DRP)



(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

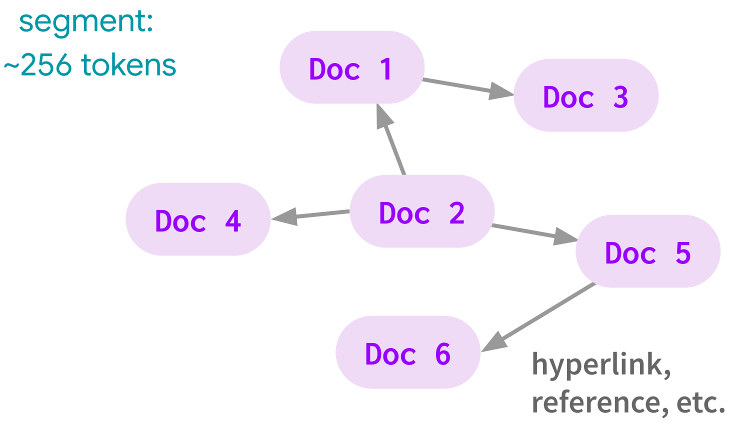


Corpus of linked documents

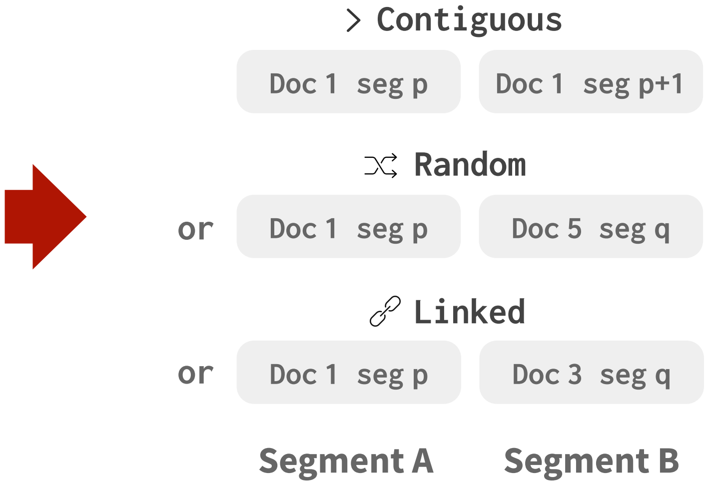
(1) Link-aware LM Input Creation

Idea

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



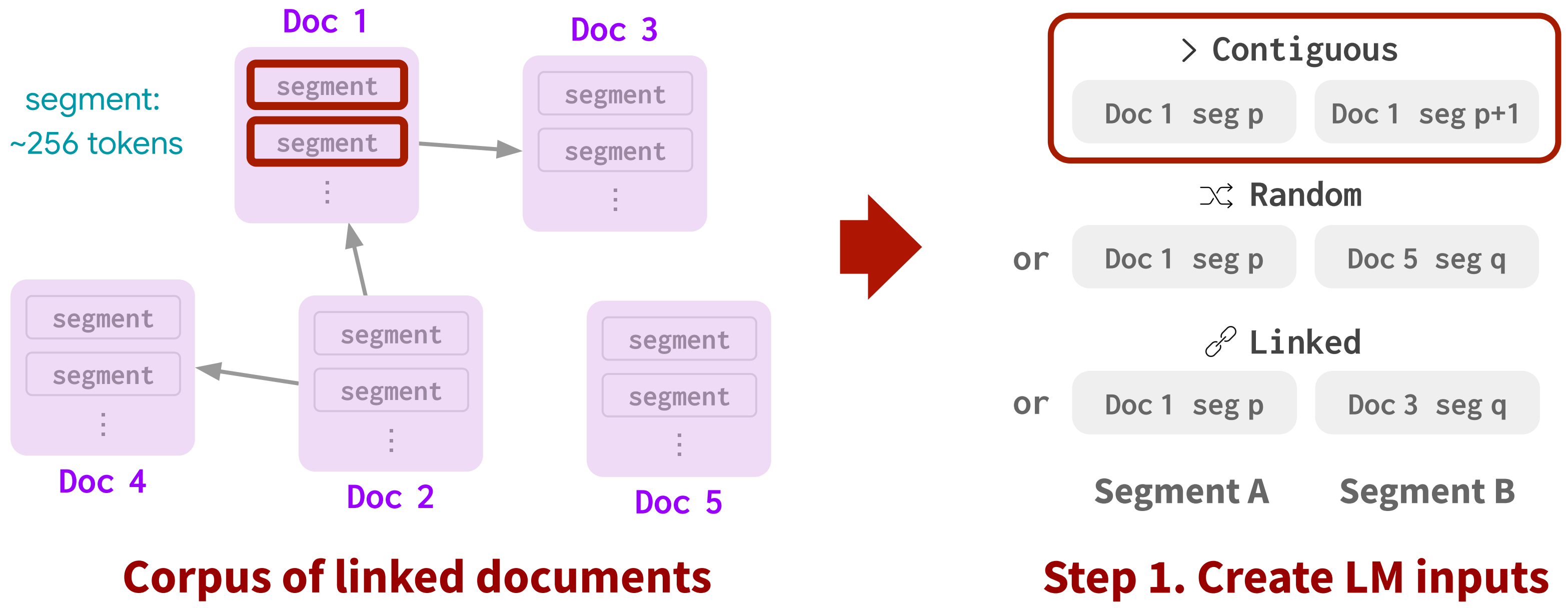
Corpus of linked documents



Step 1. Create LM inputs

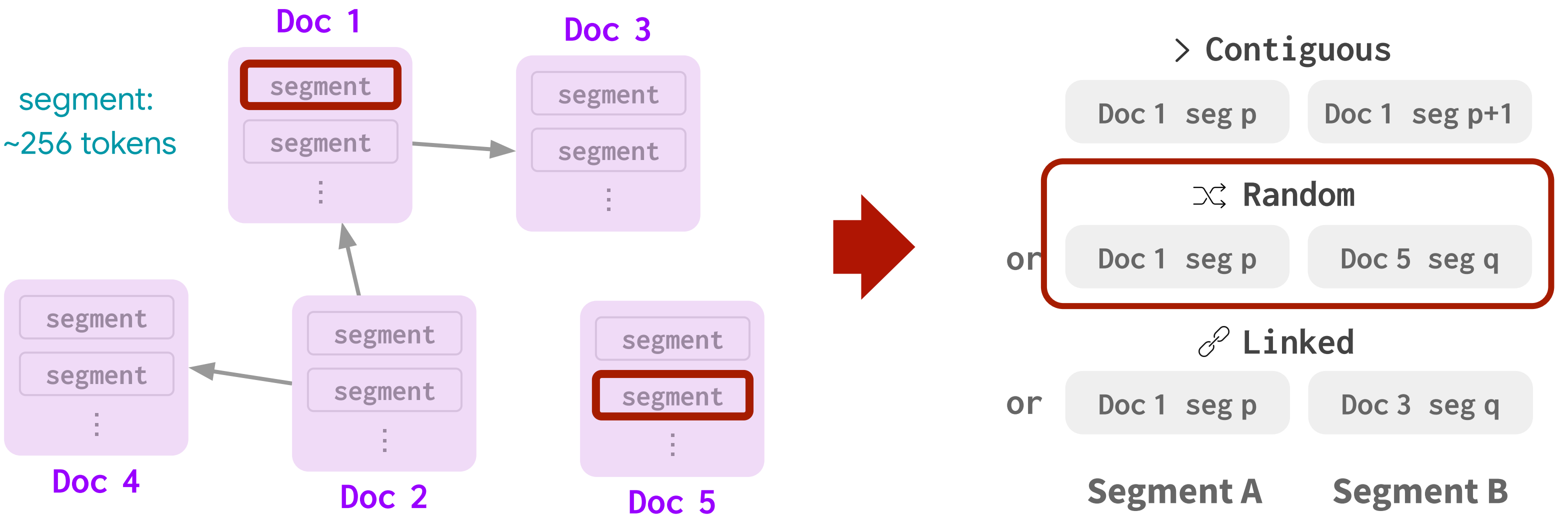
LM Input Option (i): "Contiguous"

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



LM Input Option (ii): “Random”

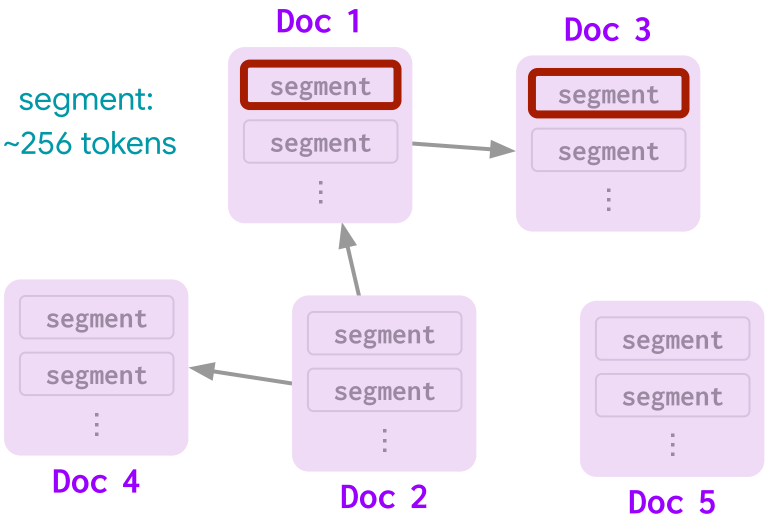
After sampling segment **A**, sample a segment from a random doc as **B** (same as BERT)



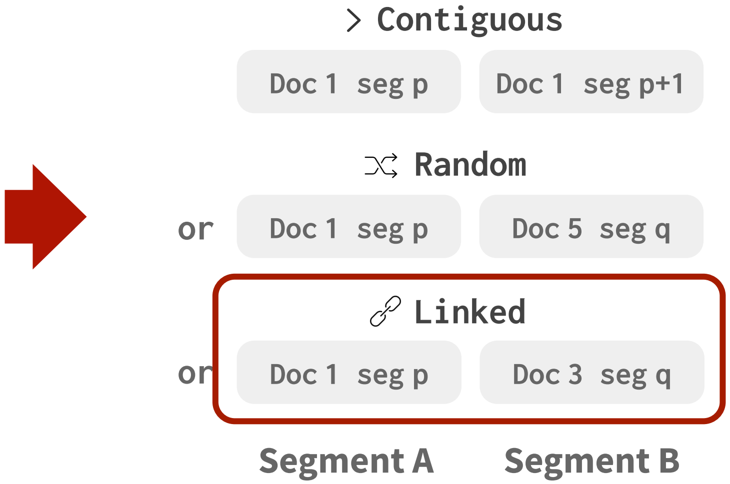
Corpus of linked documents **Step 1. Create LM inputs**

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

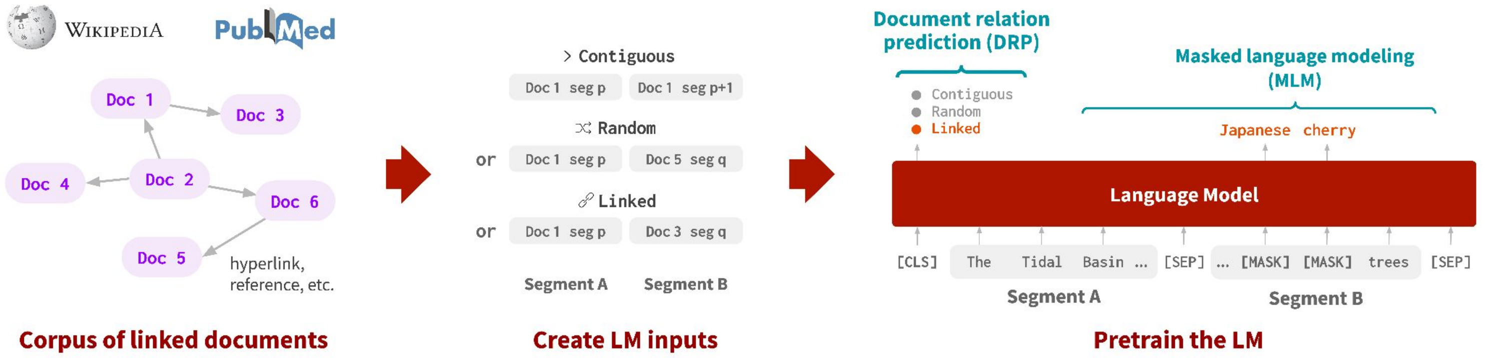
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

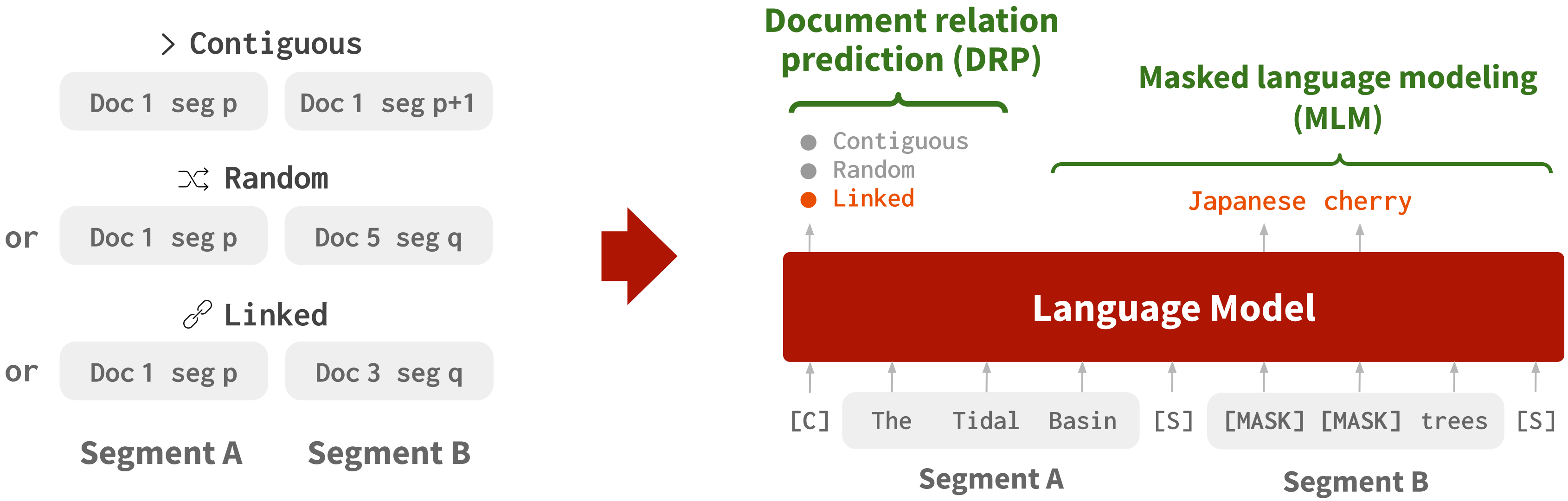
(2) Link-aware LM pretraining

- 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

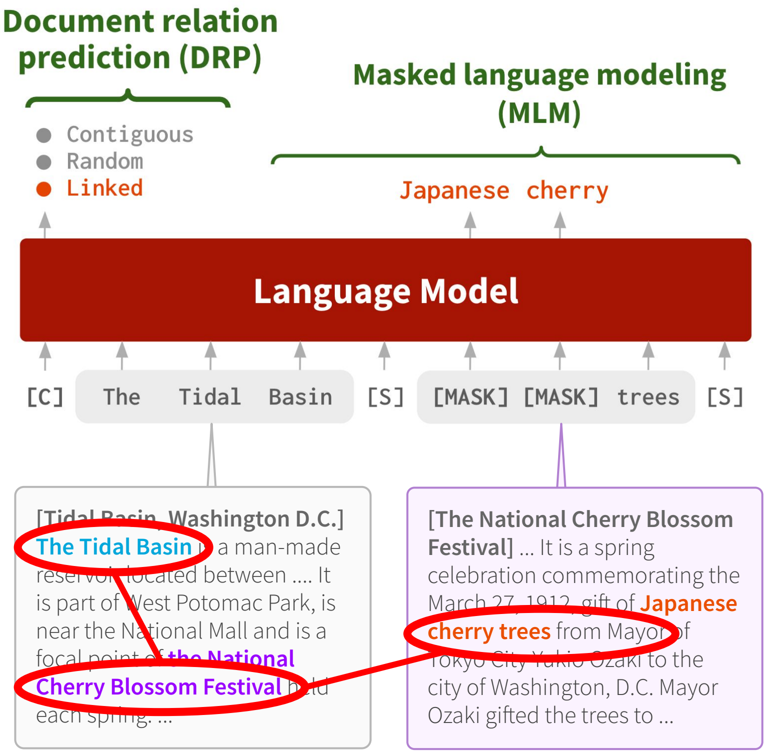
Masked language modeling (MLM)

- 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

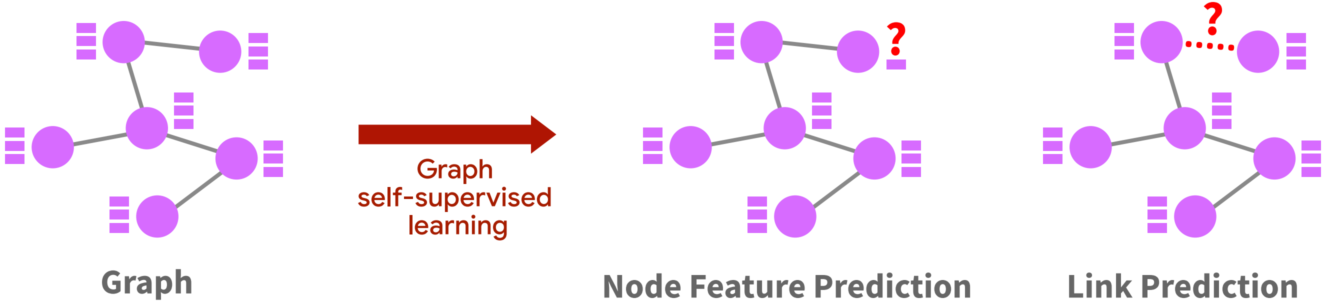
Interpretation as graph self-supervised learning on the doc graph

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

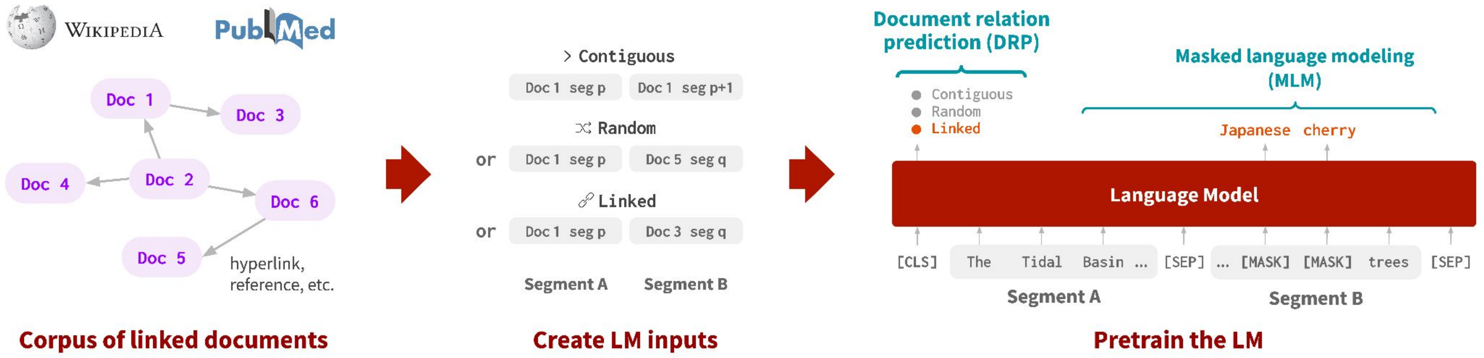
Proposed Idea: LinkBERT

(0) Document graph construction

(1) Link-aware LM input creation

(2) Link-aware LM pretraining

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



Strategy for Obtaining Linked Docs

Key factors to consider:

Relevance

The link should capture relevance. Otherwise LinkBERT is the same as BERT

⇒ ✅ Hyperlink ✅ Lexical similarity

Salience

The link should offer **new knowledge** not obvious to the current LM

⇒ ✅ Hyperlink ⚠️ Lexical similarity

Diversity

High in-degree docs may get sampled too often (e.g. “United States” page)

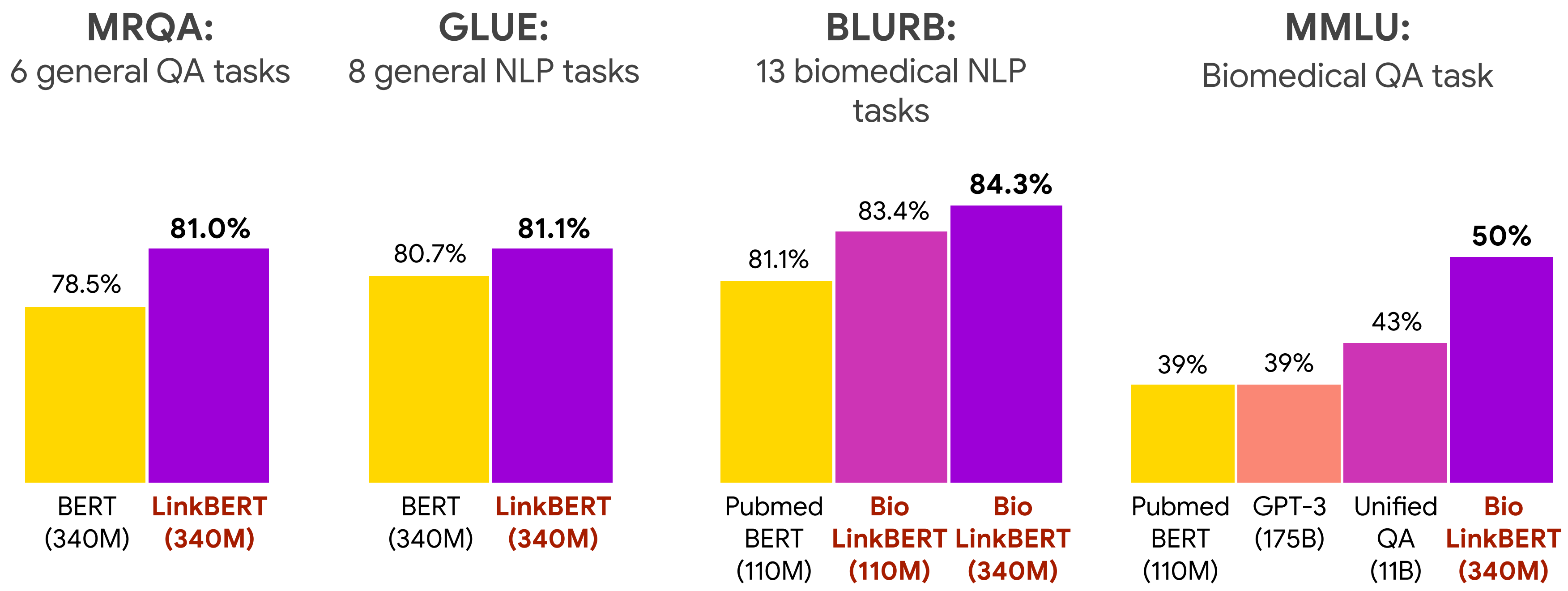
⇒ ✅ Sample a linked doc with probability inversely proportional to in-degree (Henzinger+2000)

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.

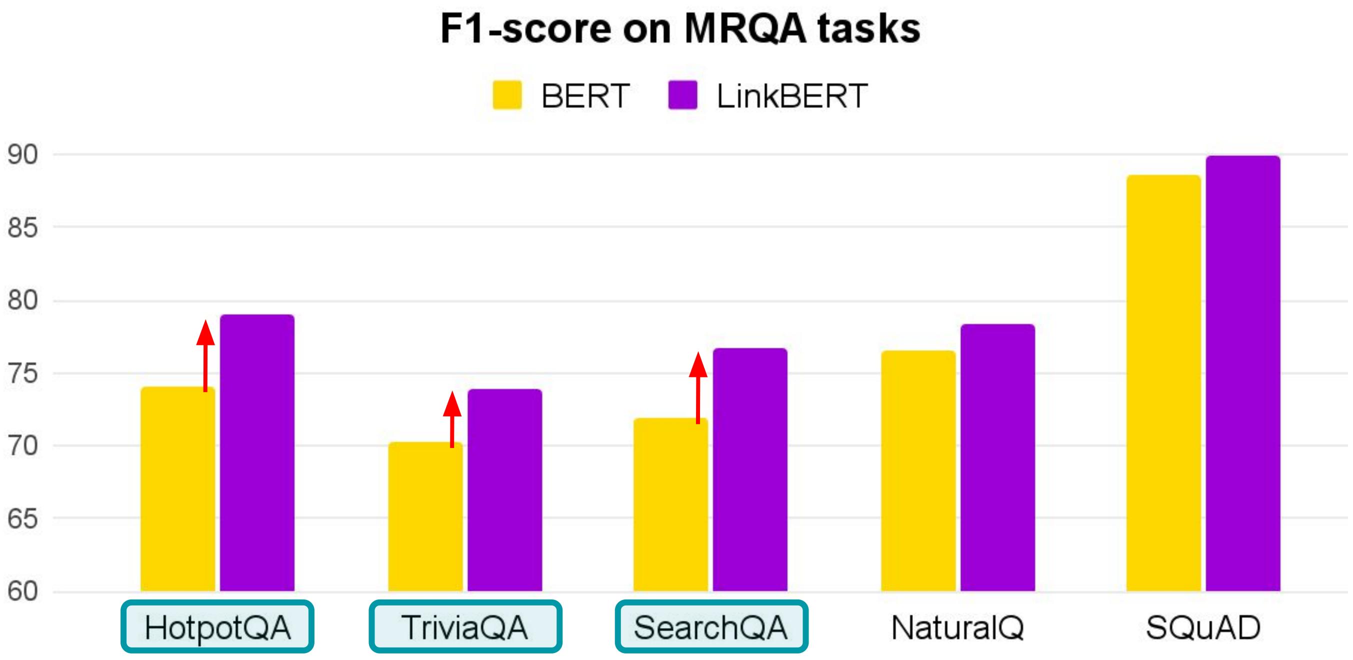
Show 100 entries

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 1: Multi-hop Reasoning

USMLE example

Question

Three days after undergoing a laparoscopic Whipple's procedure, a 43-year-old woman has swelling of her right leg. ... She was diagnosed with pancreatic cancer 1 month ago. ... Her temperature is 38°C (100.4°F), ... Which of the following is the most appropriate next step in management?

(A) CT pulmonary angiography
(B) Compression ultrasonography
(C) 2 sets of blood cultures

Need multi-hop reasoning

Leg swelling, pancreatic cancer
(symptom)

Deep vein thrombosis
(possible cause)

Compression ultrasonography
(next step for diagnosis)

Knowledge learned via document links

Doc A: ... **Pancreatic cancer** can induce **deep vein thrombosis** in leg ...
(*e.g.* Ansari et al. 2015)

Reference

Doc B: ... **Deep vein thrombosis** is tested by **compression ultrasonography** ...
(*e.g.* Piovella et al. 2002)

LinkBERT predicts: B (✓) **PubmedBERT predicts: C (✗)**

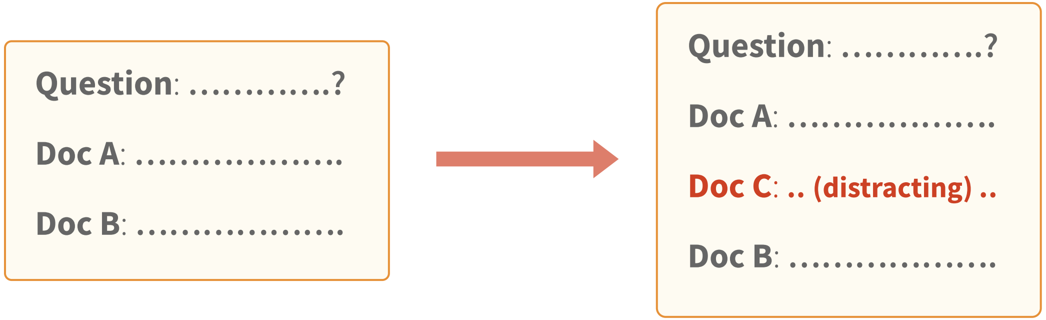
Benefit 2: Document Relation Understanding

Motivation

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

Evaluation

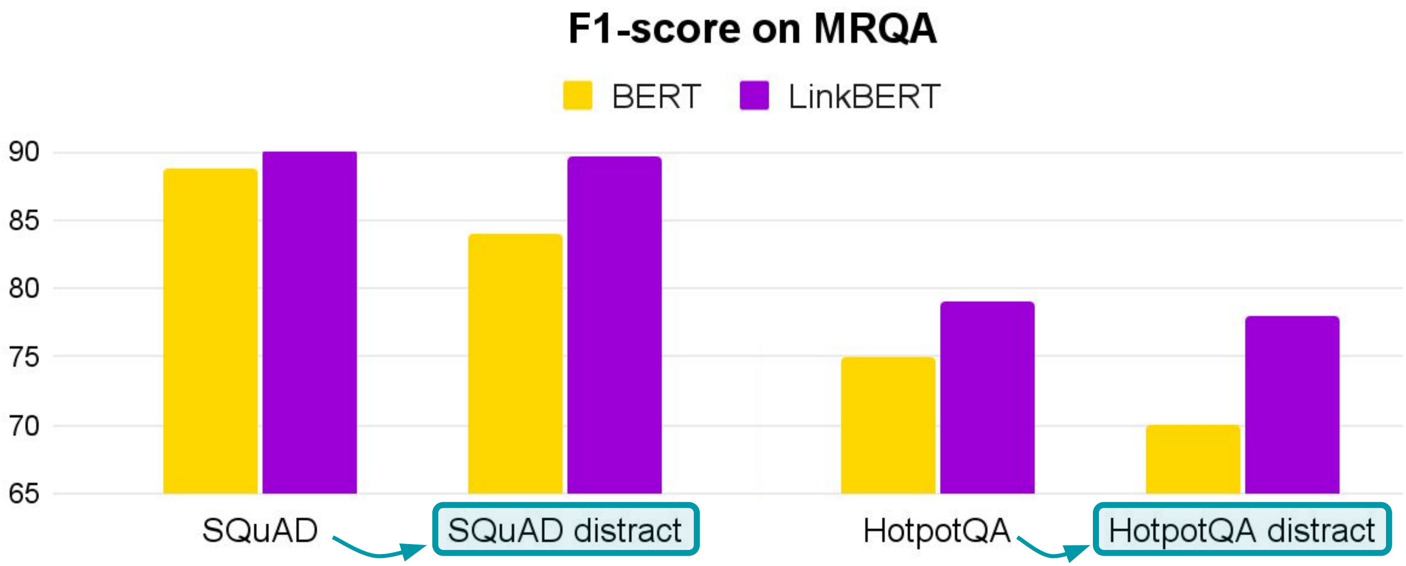
- Add distracting documents to the original MRQA datasets.
  Can LinkBERT still answer correctly?



Benefit 2: Document Relation Understanding

LinkBERT is robust to irrelevant documents

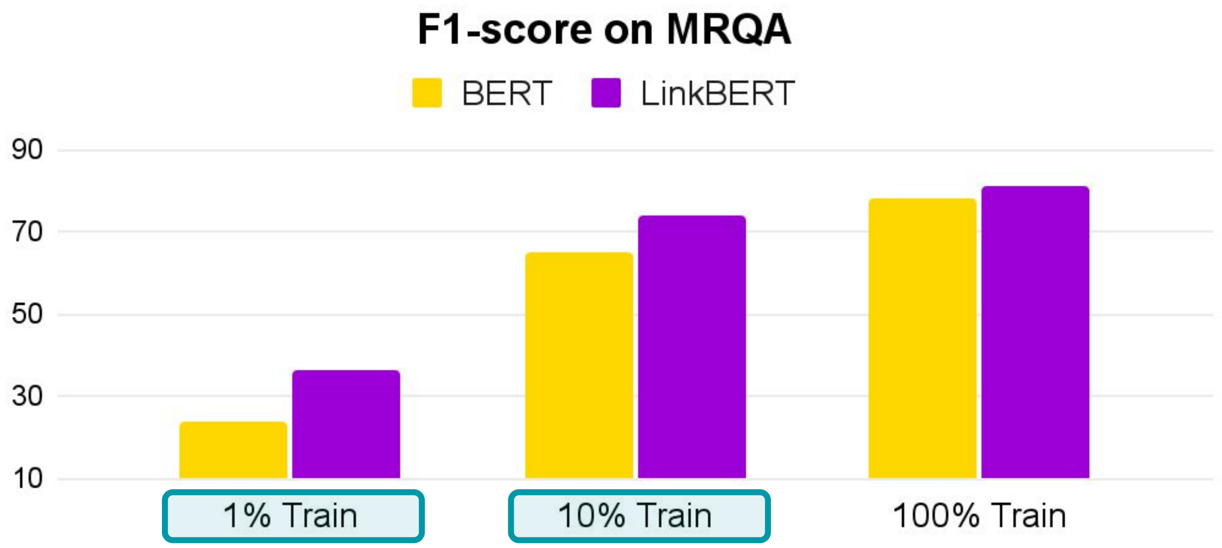
⇒ DRP task in pretraining helps recognizing doc relevance in downstream



Benefit 3: Few-shot QA

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

⇒ LinkBERT internalized more knowledge during pretraining

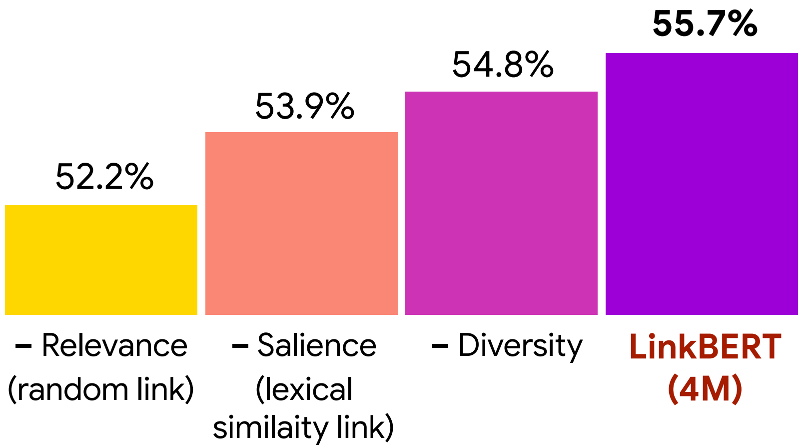


Ablation Study

Key factors for obtaining linked docs

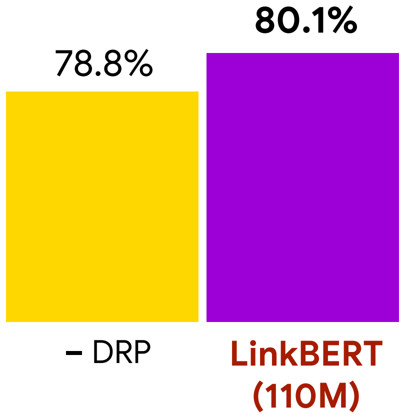
(relevance, salience, diversity)

F1-score on MRQA



Effect of DRP task in pretraining

F1-score on MRQA



Takeaways

LinkBERT: train knowledgeable LMs via document links (hyperlinks, citations)

- Place linked documents in the same LM context
- Train with joint objectives: masked LM and doc relation prediction

Benefits

- Better captures document/concept relations

  ⇒ Effective for **multi-hop** reasoning and **cross-document** understanding

- Internalizes more world knowledge

  ⇒ Effective for **knowledge-intensive** tasks, including few-shot QA

Thanks!

Michihiro Yasunaga

Percy Liang

Jure Leskovec

@michiyasunaga @percyliang @jure

Thank you to the members of the Stanford P-Lambda / SNAP / NLP groups, as well as our anonymous reviewers. Funded in part by a PECASE award, DARPA MCS, and Funai Foundation Scholarship.

Paper: LinkBERT: Pretraining Language Models with Document Links. ACL 2022.

Code/Data/Model: https://github.com/michiyasunaga/LinkBERT

HuggingFace: https://huggingface.co/michiyasunaga