LinkBERT: Pretraining Language Models with Document Links

ACL 2022

Michihiro Yasunaga, Jure Leskovec*, Percy Liang*
Stanford University
Language Model (LM) Pretraining

Core component of today’s NLP systems

Text corpus  \rightarrow  (Self-supervised) Training  \rightarrow  Pretrained LM

Complete Wikipedia and 11,038 books

Pretrained LM  \rightarrow  Adaptation

Tasks
- Question Answering
- Text Classification
- Information Retrieval
## Language Model (LM) Pretraining

Large-scale self-supervised learning

<table>
<thead>
<tr>
<th>Task</th>
<th>Examples</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masked LM</td>
<td>BERT, RoBERTa, etc.</td>
<td>‘My dog is fetching the’</td>
<td>next_word = ‘ball’</td>
</tr>
<tr>
<td>Causal LM</td>
<td>GPT-*</td>
<td>‘My __ is fetching the ball’</td>
<td>mask = ‘dog’</td>
</tr>
<tr>
<td>Seq2seq</td>
<td>BART, T5, etc.</td>
<td>‘My __ is fetching the ball’</td>
<td>denoised = ‘My dog is fetching the ball’</td>
</tr>
</tbody>
</table>
LMs learn various knowledge

https://demo.allennlp.org/nextptoken-lm
Existing LM Pretraining Methods

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

Text corpus

Pretrain the LM

Language Model

Devlin+2019, Liu+2019
But documents have rich dependencies

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

**Web:** hyperlinks

**Literature:** citations

**Code:** dependencies

---

[Diagram showing connections between documents and sources like Wikipedia, PubMed, and GitHub]
Knowledge can span across documents

Multi-hop knowledge
(e.g. Tidal Basin has Japanese cherry trees)

The Tidal Basin is a man-made reservoir located between ... It 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 ...

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](https://example.com)). 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

> Contiguous

\[ \text{Doc 1 seg } p \quad \text{Doc 1 seg } p+1 \]

> Random

\[ \text{Doc 1 seg } p \quad \text{Doc 5 seg } q \]

> Linked

\[ \text{Doc 1 seg } p \quad \text{Doc 3 seg } q \]

Segment A \quad Segment B

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)
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**

![Diagram showing the creation of LM inputs and pretraining process](Image)
(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 knowledge

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

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 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 ...
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)
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](https://example.com))
## Experiments

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

LinkBERT makes consistent improvement across tasks and domains

**MRQA:**
6 general QA tasks

- BERT (340M) 78.5%
- LinkBERT (340M) 81.0%

**GLUE:**
8 general NLP tasks

- BERT (340M) 80.7%
- LinkBERT (340M) 81.1%

**BLURB:**
13 biomedical NLP tasks

- Pubmed BERT (110M) 81.1%
- Bio LinkBERT (110M) 83.4%
- Bio LinkBERT (340M) 84.3%

**MMLU:**
Biomedical QA task

- Pubmed BERT (110M) 39%
- GPT-3 (175B) 39%
- Unified QA (11B) 43%
- Bio LinkBERT (340M) 50%
BioLinkBERT sets a new state of the art

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>BLURB Score (Macro Avg.)</th>
<th>Micro Avg.</th>
<th>NER</th>
<th>PICO</th>
<th>RE</th>
<th>SS</th>
<th>Class.</th>
<th>QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BioLinkBERT-Large — Stanford</td>
<td>84.30</td>
<td>84.80</td>
<td>86.89</td>
<td>74.19</td>
<td>82.74</td>
<td>93.63</td>
<td>84.88</td>
<td>83.50</td>
</tr>
<tr>
<td>2</td>
<td>BioLinkBERT-Base — Stanford</td>
<td>83.39</td>
<td>83.84</td>
<td>86.39</td>
<td>73.97</td>
<td>81.56</td>
<td>93.27</td>
<td>84.35</td>
<td>80.81</td>
</tr>
<tr>
<td>3</td>
<td>PubMedBERT-LARGE (fine-tuning stabilization; uncased; abstracts) — Microsoft Research</td>
<td>82.91</td>
<td>83.58</td>
<td>86.28</td>
<td>73.61</td>
<td>81.77</td>
<td>92.73</td>
<td>82.70</td>
<td>80.37</td>
</tr>
</tbody>
</table>
Benefit 1: Multi-hop Reasoning

Large gains over BERT on tasks involving multi-hop reasoning

F1-score on MRQA tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>BERT</th>
<th>LinkBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>HotpotQA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TriviaQA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SearchQA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NaturalQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQuAD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
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

Knowledge learned via document links

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

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

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

Leg swelling, pancreatic cancer (symptom)

Deep vein thrombosis (possible cause)

Compression ultrasonography (next step for diagnosis)
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?

Question: ............?
Doc A: ..................
Doc B: ..................

Question: ............?
Doc A: ..................
Doc C: .. (distracting) ..
Doc B: ..................
Benefit 2: Document Relation Understanding

LinkBERT is robust to irrelevant documents

⇒ DRP task in pretraining helps recognizing doc relevance in downstream

F1-score on MRQA

<table>
<thead>
<tr>
<th>Set</th>
<th>BERT</th>
<th>LinkBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>SQuAD distract</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>HotpotQA</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>HotpotQA distract</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>
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)

Effect of DRP task in pretraining

F1-score on MRQA

- Relevance (random link) 52.2%
- Salience (lexical similarity link) 53.9%
- Diversity 54.8%
- LinkBERT (4M) 55.7%

F1-score on MRQA

- DRP 78.8%
- LinkBERT (110M) 80.1%
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
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](https://github.com/michiyasunaga/LinkBERT)
**HuggingFace:** [https://huggingface.co/michiyasunaga](https://huggingface.co/michiyasunaga)