Empowering Language Models with Graph Learning

Michihiro Yasunaga

Joint work with Antoine Bosselut, Hongyu Ren, Xikun Zhang, Chris Manning, Percy Liang, Jure Leskovec
What is Natural Language Processing (NLP)?

- Automated understanding of natural language input
- Coherent generation of natural language output
NLP Applications

Machine Translation:

Question Answering:

Personal Assistants:

Specialized Applications:

Legal Documents
Health Records
Business Intelligence
Customer Research
Modern NLP – Powered by large language models

The New York Times
A Breakthrough for A.I. Technology: Passing an 8th-Grade Science Test

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

Unstructured text → (Self-supervised) Training → Language model → Complete Wikipedia and 11,038 books → Adaptation → Tasks: Question Answering, Text Classification, Information Retrieval...
Existing LM Pretraining

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

Text corpus

Pretrain the LM

Language Model

[iroh goes to brew [MASK] [MASK] . [s]]

[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, ....

[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 ...
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, ....

[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 ...

Language model: fine-grained local relations

Graph can provide global relations

Hyperlink

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

Linked document
Graph can bring relevant concepts closer

Text

If it is not used for **hair**, a **round brush** can be an example of what?

Knowledge Graph

Graph can provide latent relations not mentioned in text
This talk

**LinkBERT**

**DRAGON**

**General principle:** graphs bring relevant documents/concepts closer together
This talk

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

Doc 1

Doc 2

...

Doc N

Doc 1 → Doc 3

Doc 4 → Doc 2

Doc 5

Doc 6

hyperlink, etc.
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, ....

[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 ...
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 \cite{levine2022}. 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

- **Contiguous**
  - Doc 1 seg p
  - Doc 1 seg p+1

- **Random**
  - Doc 1 seg p
  - Doc 5 seg q

- **Linked**
  - Doc 1 seg p
  - Doc 3 seg q

**Step 1. Create LM inputs**

**Step 2. Pretrain the LM**

**Document relation prediction (DRP)**

**Masked language modeling (MLM)**

Language Model

- Contiguous
- Random
- Linked

Japanese cherry
(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
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

<table>
<thead>
<tr>
<th>Pretraining corpus</th>
<th>General domain</th>
<th>Biomedical domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Wikipedia (10GB) + Books (4GB)</strong></td>
<td><strong>PubMed (20GB)</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Links:</strong> hyperlinks</td>
<td><strong>Links:</strong> citations</td>
</tr>
<tr>
<td></td>
<td><strong>Doc graph:</strong> 3M nodes, 60M edges</td>
<td><strong>Doc graph:</strong> 15M nodes, 120M edges</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline</th>
<th>General domain</th>
<th>Biomedical domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>= Pretrained on same corpus, but no doc links</td>
<td><strong>BERT (Devlin+2019)</strong></td>
<td><strong>PubmedBERT (Gu+2020)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Downstream tasks</th>
<th>General domain</th>
<th>Biomedical domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GLUE</strong> (NLP benchmark)</td>
<td></td>
<td><strong>BLURB</strong> (NLP benchmark)</td>
</tr>
<tr>
<td><strong>MRQA</strong> (QA benchmark)</td>
<td></td>
<td><strong>MedQA-USMLE</strong> (QA task)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>MMLU medicine</strong> (QA task)</td>
</tr>
</tbody>
</table>
Performance

LinkBERT makes consistent improvement across tasks and domains

**MRQA:**
6 general QA tasks

![Graph showing MRQA results: BERT (340M) vs. LinkBERT (340M)]

**GLUE:**
8 general NLP tasks

![Graph showing GLUE results: BERT (340M) vs. LinkBERT (340M)]

**BLURB:**
13 biomedical NLP tasks

![Graph showing BLURB results: Pubmed BERT (110M) vs. Bio LinkBERT (340M)]

**MMLU:**
Biomedical QA task

![Graph showing MMLU results: Pubmed BERT (110M) vs. GPT-3 (175B) vs. Unified QA (11B) vs. Bio LinkBERT (340M)]
BioLinkBERT sets a new state of the art

<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 —</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></td>
<td>Stanford</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BioLinkBERT-Base —</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></td>
<td>Stanford</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>

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

Show 100 entries

https://microsoft.github.io/BLURB/leaderboard.html
Benefit 1: Multi-hop Reasoning

Large gains over BERT on tasks involving multi-hop reasoning

F1-score on MRQA tasks

- BERT
- LinkBERT

<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
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>Dataset</th>
<th>BERT</th>
<th>LinkBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>90</td>
<td>85</td>
</tr>
<tr>
<td>SQuAD distract</td>
<td>85</td>
<td>LinkBERT</td>
</tr>
<tr>
<td>HotpotQA</td>
<td>75</td>
<td>70</td>
</tr>
<tr>
<td>HotpotQA distract</td>
<td>70</td>
<td>65 LinkBERT</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
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:

```python
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

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

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

Joint Pretraining

Language-Knowledge Model
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, WQLM, KEPLER)
Proposed Method: DRAGON

Raw data

Text corpus

Knowledge graph

Self-supervised Objective

Cross-modal Encoder

Pretrain

[INT] If it is not used for hair, a round brush is an example of art supplies.

Text

Local KG

Masked LM

art supplies

LM Head

KG link prediction

(round brush, at, hair)

Fusion Layer

Interaction node

LM Layer

[INT] If it is not used for hair, a round brush is an example of [MASK].

KG Retrieval
Proposed Method: DRAGON

Step (1)
Text corpus → Knowledge graph
Text → Local KG

Step (2)
Raw data → Cross-modal Encoder

Step (3)
Self-supervised Objective
Masked LM
art supplies
LM Head
KG link prediction
(round brush, at, hair)
LinkPred Head

[Int] If it is not used for hair, a round brush is an example of art supplies.
(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
Idea: Joint self-supervised objectives

Masked LM

LM Head

Joint training
Text & KG mutually inform each other

KG Link Prediction

positive / negative link

[3] Bidirectional Self-Supervision
Proposed Method: DRAGON

**Text corpus**

**Knowledge graph**

**Self-supervised Objective**

**Cross-modal Encoder**

**Local KG**

**Pretrain**

**Masked LM**

**KG link prediction**

**LM Head**

**LinkPred Head**

[Int] If it is not used for hair, a round brush is an example of art supplies.

[Int] If it is not used for hair, a round brush is an example of [MASK] [MASK].
# Experiments

<table>
<thead>
<tr>
<th></th>
<th>General domain</th>
<th>Biomedical domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pretraining data</strong></td>
<td><strong>Text:</strong> <a href="#">BookCorpus</a> (6GB)</td>
<td><strong>Text:</strong> <a href="#">PubMed</a> (20GB)</td>
</tr>
<tr>
<td></td>
<td><strong>KG:</strong> <a href="#">ConceptNet</a> (800K nodes, 2M edges)</td>
<td><strong>KG:</strong> <a href="#">UMLS</a> (300K nodes, 1M edges)</td>
</tr>
<tr>
<td><strong>Downstream tasks</strong></td>
<td>Commonsense reasoning (OBQA, RiddleSense, CommonsenseQA, CosmosQA, HellaSwag, PIQA, SIQA, aNLI, ARC)</td>
<td>Biomedical reasoning (PubMedQA, BioASQ, MedQA-USMLE)</td>
</tr>
<tr>
<td><strong>Baseline: LM</strong></td>
<td>RoBERTa (<a href="#">Liu+2019</a>)</td>
<td>BioLinkBERT (<a href="#">Yasunaga+2022</a>)</td>
</tr>
</tbody>
</table>

**Ours (DRAGON): LM pretrained with KG**
Performance

DRAGON makes consistent improvement across tasks and domains

**Commonsense reasoning tasks**
(e.g. OBQA, RiddleSense)

<table>
<thead>
<tr>
<th>Model</th>
<th>Effect of Pretraining</th>
<th>Effect of KG</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td>69.0%</td>
<td></td>
</tr>
<tr>
<td>+Grease LM</td>
<td>70.4%</td>
<td></td>
</tr>
<tr>
<td>DRAGON (Ours)</td>
<td>75.1%</td>
<td></td>
</tr>
</tbody>
</table>

**Biomedical reasoning tasks**
(e.g. PubMedQA, MedQA)

<table>
<thead>
<tr>
<th>Model</th>
<th>Effect of Pretraining</th>
<th>Effect of KG</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioLink BERT</td>
<td>70.5%</td>
<td></td>
</tr>
<tr>
<td>+Grease LM</td>
<td>70.6%</td>
<td></td>
</tr>
<tr>
<td>DRAGON (Ours)</td>
<td>72.5%</td>
<td></td>
</tr>
</tbody>
</table>
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

---

**DRAGON**  
**GNN 1st Layer**  
**Int**

**folding chair**  
**school**

**camp**  
**beach**

---

**DRAGON**  
**GNN Final Layer**  
**Int**

**folding chair**  
**school**

**camp**  
**beach**

---

**Model Prediction**

RoBERTa:  
A. camp (x)

GreaseLM:  
C. camp (x)

**DRAGON: B. school (✓)**

### Negation + Conjunction

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

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

---

**DRAGON**  
**GNN 1st Layer**  
**Int**

**folding chair**  
**school**

**camp**  
**beach**

---

**DRAGON**  
**GNN Final Layer**  
**Int**

**folding chair**  
**school**

**camp**  
**beach**

---

**Model Prediction**

RoBERTa:  
B. school (x)

GreaseLM:  
B. school (x)

**DRAGON: C. beach (✓)**

---

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

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. 

- Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, Jure Leskovec. 

- Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Chris Manning, Jure Leskovec. 

- Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Chris Manning, Percy Liang, Jure Leskovec. 

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