DRAGON:
Deep Bidirectional Language-Knowledge Pretraining

NeurIPS 2022

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

Text

(Self-supervised) Training

Language model

Complete Wikipedia and 11,038 books

Adaptation

Tasks

Question Answering

Text Classification

Information Retrieval

...
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)
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 Foundation Model
Proposed Method: DRAGON

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

If it is not used for hair, a round brush is an example of art supplies.
Proposed Method: DRAGON

Step (1)

Text corpus
Knowledge graph

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

Knowledge graph
Local KG

Step (2)

Cross-modal Encoder

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

Step (3)

Self-supervised Objective

Masked LM
KG link prediction

art supplies
(round brush, at, hair)

LM Layer

KG Retrieval

Step (4)

Pretrain
(1) Text-KG Input

Motivation
● Informative pair of (text, local KG):
  Text can contextualize the KG
  KG can ground the text

Idea
● Given text corpus and KG, sample a text segment and retrieve a relevant knowledge subgraph by entity linking
  ⇒ Aligned pairs of (text, local KG)
**Idea**

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

(2) Deep Bidirectional Cross-Modal Model

---

**Zhang et al. 2022**
(3) Bidirectional Self-Supervision

Idea: Pretrain with two self-supervised reasoning tasks

Masked LM

LM Head

hair

round brush is used for [MASK]

KG Link Prediction

LinkPred Head

positive / negative link

Joint training

Text & KG mutually inform each other

RelatedTo

UsedFor

AtLocation?

AtLocation

Round brush

Paint

Art supply

Hairbrush
Proposed Method: DRAGON

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

Text corpus

Knowledge graph

Self-supervised Objective

Cross-modal Encoder

Local KG

Masked LM

KG link prediction

Pretrain

Text

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

Ours (DRAGON): LM pretrained with KG
DRAGON makes consistent improvement across tasks and domains

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

- RoBERTa + Grease LM: 69.0%
- DRAGON (Ours): 70.4%
- Effect of pretraining:
  - RoBERTa: 69.0%
  - DRAGON (Ours): 75.1%

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

- BioLink BERT + Grease LM: 70.5%
- DRAGON (Ours): 72.5%
- Effect of pretraining:
  - BioLink BERT: 70.5%
  - DRAGON (Ours): 72.5%

Effect of KG:
- DRAGON (Ours): 75.1%
- DRAGON (Ours): 72.5%
Benefit 1: Complex Reasoning

Large gains on QA examples involving complex reasoning

RoBERTa  GreaseLM  DRAGON

Negation  Conjunction  Prepositional Phrases > 2  Entities > 10  Other Questions
Benefit 1: Complex Reasoning

**Conjunction**

Where would you use a *folding chair* and store one?
- A. camp
- B. *school*
- C. beach

**Negation + Conjunction**

Where would you use a *folding chair* but not store one?
- A. garage
- B. *school*
- C. beach

In DRAGON, KG serves as *scaffold* for performing *structured reasoning*.
Benefit 1: Complex Reasoning

**Single context**

You will buy a ticket for entering what building for entertainment?

A. station  
B. movie theater

**Multi context (extra reasoning step)**

You don't enjoy watching pre-recorded performance. You will buy a ticket for entering what building for entertainment?

A. station  
B. movie theater  
C. concert hall

---

**DRAGON GNN 1st Layer**

**DRAGON GNN Final Layer**

**Model Prediction**

---

RoBERTa:  
B. movie theater (✓)

GreaseLM:  
B. movie theater (✓)

DRAGON:  
B. movie theater (✓)

---

RoBERTa:  
B. movie theater (✗)

GreaseLM:  
B. movie theater (✗)

DRAGON:  
C. concert hall (✓)

---

Pretraining with KG helps **extrapolate** to harder test examples that need **multi-step reasoning**.
Benefit 2: Low-Resource QA

Large gains on few-shot and low-resource QA

⇒ Intuition: self-supervision helps learn more knowledge
Key Design Choices: Modeling

Cross-modal fusion for text+KG

- Bidirectional interaction (DRAGON)
- Concatenate representations at end

Accuracy on OBQA

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concatenate at end</td>
<td>68.0%</td>
</tr>
<tr>
<td>Bidirectional interaction (DRAGON)</td>
<td>72.0%</td>
</tr>
</tbody>
</table>

KG structure

- Use graph and GNN (DRAGON)
- Convert to sentence and add to text

Accuracy on OBQA

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convert to sentence</td>
<td>70.1%</td>
</tr>
<tr>
<td>Use graph (DRAGON)</td>
<td>72.0%</td>
</tr>
</tbody>
</table>
Key Design Choices: Self-Supervision

**Pretraining objective**
- Joint MLM + LinkPred (DRAGON)
- MLM only
- LinkPred only

**LinkPred head**
- DistMult (Final DRAGON)
- TransE
- RotatE

⇒ All help

**Accuracy on OBQA**
- LinkPred only: 66.4%
- MLM only: 67.2%
- MLM + LinkPred (DRAGON): 72.0%

**Accuracy on OBQA**
- No LinkPred: 67.2%
- TransE: 71.4%
- RotatE: 71.7%
- DistMult (DRAGON): 72.0%
**DRAGON**: Pretrain a foundation model jointly on text & KG

**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)
We thank the members of the Stanford SNAP / P-Lambda / NLP groups, and the project MOWGLI team, as well as our anonymous reviewers. Funded in part by DARPA MCS.

**Paper:** Deep Bidirectional Language-Knowledge Graph Pretraining. NeurIPS 2022.

**Code:** [https://github.com/michiyasunaga/DRAGON](https://github.com/michiyasunaga/DRAGON)