Knowledge-Augmented PreTraining for Reasoning

ACL 2023

Tutorial on Complex Reasoning in Natural Language

Michihiro Yasunaga
Stanford University
Recap: Knowledge

Knowledge is available in various forms

Text

- Diverse & contextual knowledge

Knowledge Graph (KG)

- Structured background knowledge
Knowledge helps complex reasoning

Reasoning often involves combining multiple pieces of knowledge

Lin+2019; Yasunaga+2021
This section: Knowledge-augmented Pre-Training
What is Pre-Training (and Pre-Finetuning)

- **Key**: learn from diverse data (e.g. through self-supervised learning)

Wide range of tasks:
- Question Answering
- Text Classification
- Information Retrieval

Diverse data → Model

Pre-train
(e.g. language modeling, self-supervised learning)

Pre-finetune
(e.g. instruction tuning, RLHF)

Adapt
(e.g. finetuning, prompting)
Why Pre-Training (and Pre-Finetuning)?

- Help a broad range of downstream tasks
- Make adaptation efficient (e.g. few-shot finetuning/prompting)

Diverse data → Model

Pre-train
(e.g. language modeling, self-supervised learning)

Pre-finetune
(e.g. instruction tuning, RLHF)

Adapt
(e.g. finetuning, prompting)

Wide range of tasks

- Question Answering
- Text Classification
- Information Retrieval

...
Goal: Knowledge-augmented Pre-Training

Text

- Diverse & contextual knowledge

Knowledge-augmented Pre-training

Knowledge Graph (KG)

- Structured background knowledge

Knowledge- & Reasoning-intensive Tasks
# Outline of Knowledge-augmented Pre-training

## Integrate textual knowledge

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>REALM</strong> [ICML 2020]</td>
<td><strong>CDLM</strong> [EMNLP 2021]</td>
<td><strong>LinkBERT</strong> [ACL 2022]</td>
</tr>
</tbody>
</table>

## Integrate structured knowledge

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WKLM</strong> [ICLR 2020]</td>
<td><strong>KEPLER</strong> [TACL 2021]</td>
<td><strong>JAKET</strong> [AAAI 2022]</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ERNIE</strong> [ACL 2019]</td>
<td><strong>CoLAKE</strong> [COLING 2020]</td>
<td><strong>DRAGON</strong> [NeurIPS 2022]</td>
</tr>
</tbody>
</table>
# Integrate Textual Knowledge

## Integrate Textual Knowledge

<table>
<thead>
<tr>
<th>Integrate structured knowledge</th>
<th>Knowledge graph as training objective</th>
<th>Knowledge graph as input context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>REALM</strong> [ICML 2020]</td>
<td><strong>ERNIE</strong> [ACL 2019]</td>
</tr>
<tr>
<td></td>
<td><strong>CDLM</strong> [EMNLP 2021]</td>
<td><strong>CoLAKE</strong> [COLING 2020]</td>
</tr>
<tr>
<td></td>
<td><strong>LinkBERT</strong> [ACL 2022]</td>
<td><strong>DRAGON</strong> [NeurIPS 2022]</td>
</tr>
<tr>
<td></td>
<td><strong>WKLM</strong> [ICLR 2020]</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>KEPLER</strong> [TACL 2021]</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>JAKET</strong> [AAAI 2022]</td>
<td></td>
</tr>
</tbody>
</table>
Integrate Textual Knowledge

- We will focus on **text retrieval**, which helps to make reasoning process more explicit.
REALM: Retrieval-Augmented Language Model Pre-Training

Method

- When doing masked token prediction, **retrieve** relevant documents from a knowledge corpus as reference.

Key

- **DPR** was retrieval-augmented finetuning for QA. REALM is a **self-supervised pre-training** version.

REALM: Retrieval-Augmented Language Model Pre-Training

Result

- Improve knowledge-intensive NLP (e.g. open-domain QA)

Figure from https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1224/slides/cs224n-2022-lecture15-guu.pdf
CDLM: Cross-Document Language Modeling

Method
- Retrieve related docs and pre-train LM on concatenated context
- Internalize knowledge during pretraining. Retrieval is optional during inference.

Set of related docs
- Doc 1: “Harry Shearer is suing Vivendi’s Universal Music for $125 million for allegedly fraudulent…”
- Doc 2: “...Harry Shearer alleges parent company of Universal Music and StudioCanal withheld millions…”
- Doc 3: “Shearer was then joined in the lawsuit with StudioCanal and its French parent Vivendi by his co-stars”

LM pre-training

CDLM: Cross-Document Language Modeling

Result

- Improve cross-document NLP (e.g. citation recommendation, coreference resolution)

Takeaway

- Retrieval-augmented pre-training helps cross-document reasoning

(Citation recommendation illustration: [link](https://arxiv.org/pdf/2108.07571.pdf))

<table>
<thead>
<tr>
<th>Model</th>
<th>AAN</th>
<th>OC</th>
<th>S2orc</th>
<th>PAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMASH (2019)</td>
<td>80.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SMITH (2020)</td>
<td>85.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BERT-HAN (2020)</td>
<td>65.0</td>
<td>86.3</td>
<td>90.8</td>
<td>87.4</td>
</tr>
<tr>
<td>GRU-HAN+CDA (2020)</td>
<td>75.1</td>
<td>89.9</td>
<td>91.6</td>
<td>78.2</td>
</tr>
<tr>
<td>BERT-HAN+CDA (2020)</td>
<td>82.1</td>
<td>87.8</td>
<td>92.1</td>
<td>86.2</td>
</tr>
<tr>
<td>Longformer</td>
<td>85.4</td>
<td>93.4</td>
<td>95.8</td>
<td>80.4</td>
</tr>
<tr>
<td>Local CDLM</td>
<td>83.8</td>
<td>92.1</td>
<td>94.5</td>
<td>80.9</td>
</tr>
<tr>
<td>Rand CDLM</td>
<td>85.7</td>
<td>93.5</td>
<td>94.6</td>
<td>79.4</td>
</tr>
<tr>
<td>Prefix CDLM</td>
<td>87.3</td>
<td>94.8</td>
<td>94.7</td>
<td>81.7</td>
</tr>
<tr>
<td>CDLM</td>
<td><strong>88.8</strong></td>
<td><strong>95.3</strong></td>
<td><strong>96.5</strong></td>
<td><strong>82.9</strong></td>
</tr>
</tbody>
</table>

Table 4: $F_1$ scores over the document matching benchmarks’ test sets.

Method

- Retrieve related docs and pre-train LM on concatenated context
- Include various doc relations (e.g. hyperlink, citation, dense retrieval)
- Internalize knowledge during pretraining. Retrieval is optional during inference.

Result

- Improve knowledge-intensive NLP
- Improve multi-hop & multi-document reasoning

LinkBERT: Pretraining Language Models with Document Links

Takeaway

- Retrieval-augmented pre-training (⇒ multi-document in context) helps learn 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” (✗)

## Summary so far

| Integrate textual knowledge | REALM - retrieve relevant docs  
|                            | CDLM - learn relevant docs  
|                            | LinkBERT - learn relevant docs & doc relations  
| Integrate structured knowledge | WKLM - KG entity objective  
| Knowledge graph as training objective | KEPLER - KG link objective  
| Knowledge graph as input context | JAKET - both entity and link objectives  
| Knowledge graph as input context | ERNIE - contextualize entity emb  
|                                | CoLAKE - contextualize KG triplet  
|                                | DRAGON - contextualize KG subgraph  

## Integrate Structured Knowledge

<table>
<thead>
<tr>
<th>Integrate textual knowledge</th>
<th>REALM [ICML 2020]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDLM [EMNLP 2021]</td>
<td>LinkBERT [ACL 2022]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Integrate structured knowledge</th>
<th>WKLM [ICLR 2020]</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEPLER [TACL 2021]</td>
<td>JAKET [AAAI 2022]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge graph as input context</th>
<th>ERNIE [ACL 2019]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLAKE [COLING 2020]</td>
<td>DRAGON [NeurIPS 2022]</td>
</tr>
</tbody>
</table>
Integrate Knowledge Graph (KG)

**KG as objective (output)**
- Convenient – KG not needed at test time

**KG as input**
- Expressive model
Integrate Knowledge Graph as Training Objective

| Integrate textual knowledge | REALM [ICML 2020]  
|                            | CDLM [EMNLP 2021]  
|                            | LinkBERT [ACL 2022] |
| Integrate structured knowledge | WKLM [ICLR 2020]  
| knowledge                  | KEPLER [TACL 2021]  
|                            | JAKET [AAAI 2022]   |
| Knowledge graph as input context | ERNIE [ACL 2019]  
|                            | CoLAKE [COLING 2020]  
|                            | DRAGON [NeurIPS 2022] |
Information in Knowledge Graph

- Entity information (e.g., category, definition)
- Link information (e.g., reasoning about entity relations)

Figure from [http://web.stanford.edu/class/cs224w/slides/10-kg.pdf](http://web.stanford.edu/class/cs224w/slides/10-kg.pdf)
WKLM: Weakly Supervised Knowledge-Pretrained Language Model

Idea

- Add entity prediction objective

Method

- Replace entity mentions in text by false entities in the same category
- Predict true/false entities

WKLM: Weakly Supervised Knowledge-Pretrained Language Model

Result
● Improve knowledge-intensive NLP (e.g. QA, entity typing)

Takeaway
● Seminal work in using KG for LM pre-training objective

<table>
<thead>
<tr>
<th>Model</th>
<th>SQuAD (F1)</th>
<th>TriviaQA (F1)</th>
<th>Quasar-T (F1)</th>
<th>FIGER (acc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WKLM</td>
<td>91.3</td>
<td>56.7</td>
<td>49.9</td>
<td>60.21</td>
</tr>
<tr>
<td>WKLM w/o MLM</td>
<td>87.6</td>
<td>52.5</td>
<td>48.1</td>
<td>58.44</td>
</tr>
<tr>
<td>BERT + 1M Updates</td>
<td>91.1</td>
<td>56.3</td>
<td>48.2</td>
<td>54.17</td>
</tr>
</tbody>
</table>

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

Idea

- Add KG **link prediction** objective

Method

- Predict whether (head, relation, tail) forms a link or not
- Use TransE head: $\| h + r - t \|$
KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

**Result**
- Improve knowledge-intensive NLP and KG link prediction

**Takeaway**
- Besides entities, KG links (structure) can augment LM pre-training objective

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERNIE&lt;sub&gt;BERT&lt;/sub&gt;</td>
<td>70.0</td>
<td>66.1</td>
<td>68.0</td>
</tr>
<tr>
<td>KnowBert&lt;sub&gt;BERT&lt;/sub&gt;</td>
<td>73.5</td>
<td>64.1</td>
<td>68.5</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>70.4</td>
<td>71.1</td>
<td>70.7</td>
</tr>
<tr>
<td>ERNIE&lt;sub&gt;RoBERTa&lt;/sub&gt;</td>
<td>73.5</td>
<td>68.0</td>
<td>70.7</td>
</tr>
<tr>
<td>KnowBert&lt;sub&gt;RoBERTa&lt;/sub&gt;</td>
<td>71.9</td>
<td>69.9</td>
<td>70.9</td>
</tr>
<tr>
<td><strong>KEPLER-Wiki</strong></td>
<td>71.5</td>
<td>72.5</td>
<td>72.0</td>
</tr>
</tbody>
</table>

Table 5: Precision, recall and F-1 on TACRED (%).

<table>
<thead>
<tr>
<th>Model</th>
<th>MR</th>
<th>MRR</th>
<th>HITS@1</th>
<th>HITS@3</th>
<th>HITS@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKRL (Xie et al., 2016)</td>
<td>78</td>
<td>23.1</td>
<td>5.9</td>
<td>32.0</td>
<td>54.6</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>723</td>
<td>7.4</td>
<td>0.7</td>
<td>1.0</td>
<td>19.6</td>
</tr>
<tr>
<td><strong>KEPLER-Wiki</strong></td>
<td>32</td>
<td>35.1</td>
<td>15.4</td>
<td>46.9</td>
<td>71.9</td>
</tr>
<tr>
<td><strong>KEPLER-Cond</strong></td>
<td>28</td>
<td>40.2</td>
<td>22.2</td>
<td>51.4</td>
<td>73.0</td>
</tr>
</tbody>
</table>

(b) Inductive results on Wikidata5M (% except MR).

JAKET: Joint Pre-training of Knowledge Graph and Language Understanding

Idea
- Add both KG entity and link prediction objectives

Result
- Further improvement on multi-hop reasoning (e.g., MetaQA)

MetaQA example (2-hop): “Who acted in the movies directed by Erik Poppe?”

<table>
<thead>
<tr>
<th>Model</th>
<th>KG-Full 1-hop</th>
<th>KG-Full 2-hop</th>
<th>KG-50% 1-hop</th>
<th>KG-50% 2-hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td>90.2</td>
<td>70.8</td>
<td>61.5</td>
<td>39.3</td>
</tr>
<tr>
<td>RoB+G+M</td>
<td>91.4</td>
<td>72.6</td>
<td>62.5</td>
<td>40.8</td>
</tr>
<tr>
<td><strong>JAKET</strong></td>
<td><strong>93.9</strong></td>
<td><strong>73.2</strong></td>
<td><strong>63.1</strong></td>
<td><strong>41.9</strong></td>
</tr>
</tbody>
</table>

Table 2: Results on the MetaQA dataset over 1-hop and 2-hop questions under KG-Full and KG-50% settings. RoB+G+M is the abbreviation for the baseline model RoBERTa+GNN+M.
## Summary so far

| Integrate textual knowledge | REALM - retrieve relevant docs  
|                            | CDLM - learn relevant docs  
|                            | LinkBERT - learn relevant docs & doc relations |
| Integrate structured knowledge | WKLM - KG entity objective  
| Knowledge graph as training objective | KEPLER - KG link objective  
|                                  | JAKET - both entity and link objectives |
| Knowledge graph as input context | ERNIE - contextualize entity emb  
|                                  | CoLAKE - contextualize KG triplet  
|                                  | DRAGON - contextualize KG subgraph |
## Integrate Knowledge Graph as Input Context

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrate structured knowledge</td>
<td>Knowledge graph as training objective</td>
<td>WKLM [ICLR 2020]</td>
<td>KEPLER [TACL 2021]</td>
</tr>
</tbody>
</table>
ERNIE: Enhanced Language Representation with Informative Entities

Method
● Add KG **entity embeddings** in LM context
● Entity embs are concatenated to corresponding word embs

Takeaway
● Seminal work in using KG info as input in LM pre-training

CoLAKE: Contextualized Language and Knowledge Embedding

Method

- Add **KG triplets** in LM context
  \((\Rightarrow \text{bring neighbor entities and relations in context, besides direct entity mentions in text})\)

- Masked token prediction for both text and KG sides

---

**Example:**

**Text**

*Harry Potter* points his wand at *Lord Voldemort*

**KG triplets**

*(Harry Potter, mother, Lily Potter),
(Harry Potter, spouse, Ginny Weasley),
(Harry Potter, enemy of, Lord Voldemort)*

---

CoLAKE: Contextualized Language and Knowledge Embedding

Result

- Improve knowledge-intensive NLP and KG link prediction

Takeaway

- KG triplets provide background knowledge and help reason about related entities

<table>
<thead>
<tr>
<th>Model</th>
<th>MR ↓</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE (Bordes et al., 2013)</td>
<td>15.97</td>
<td>67.30</td>
</tr>
<tr>
<td>DistMult (Yang et al., 2015)</td>
<td>27.09</td>
<td>60.56</td>
</tr>
<tr>
<td>ComplEx (Trouillon et al., 2016)</td>
<td>26.73</td>
<td>61.09</td>
</tr>
<tr>
<td>RotatE (Sun et al., 2019)</td>
<td>30.36</td>
<td>70.90</td>
</tr>
<tr>
<td>CoLAKE</td>
<td>2.03</td>
<td>82.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>MR ↓</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKRL (Xie et al., 2016)</td>
<td>168.21</td>
<td>8.18</td>
</tr>
<tr>
<td>CoLAKE</td>
<td>31.01</td>
<td>28.10</td>
</tr>
</tbody>
</table>

Table 5: The experimental results on word-knowledge graph completion task.

Method

- Add **KG subgraph** in input context
- Text is contextualized by LM and KG is contextualized by GNN. The two are then contextualized bidirectionally

Yasunaga, et al. "**DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining**" NeurIPS 2022
DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining

Result

- Improve broad reasoning tasks (QA, commonsense, link prediction)
- Improve complex reasoning (multi-hop, logical)

Takeaway

- KG graph structure provides LM with a scaffold to perform complex reasoning about entities.

If it is not used for hair, a round brush is an example of what?
A. hair brush  B. art supplies*

After several layers of fusion, attention weight from text over hair decreases, but attention weight over round brush and painting increases, adjusting for the negation in text.

**Takeaway**

- **KG graph structure** provides LM with a **scaffold** to perform complex reasoning about entities.

---

**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

---

Yasunaga, et al. "**DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining**" NeurIPS 2022
## Summary so far

| Integrate textual knowledge | REALM - retrieve relevant docs  
CDLM - learn relevant docs  
LinkBERT - learn relevant docs & doc relations |
|-----------------------------|------------------------------------------------------------------|
| Integrate structured knowledge | Knowledge graph as training objective  
WKLM - KG entity objective  
KEPLER - KG link objective  
JAKET - both entity and link objectives |
| Knowledge graph as input context | ERNIE - contextualize entity emb  
CoLAKE - contextualize KG triplet  
DRAGON - contextualize KG subgraph |
## Summary

<table>
<thead>
<tr>
<th>Integrate textual knowledge</th>
<th>REALM - retrieve relevant docs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CDLM - learn relevant docs</td>
</tr>
<tr>
<td></td>
<td>LinkBERT - learn relevant docs &amp; doc relations</td>
</tr>
<tr>
<td>Integrate structured knowledge</td>
<td>Knowledge graph as training objective</td>
</tr>
<tr>
<td></td>
<td>WKLM - KG entity objective</td>
</tr>
<tr>
<td></td>
<td>KEPLER - KG link objective</td>
</tr>
<tr>
<td></td>
<td>JAKET - both entity and link objectives</td>
</tr>
<tr>
<td>Knowledge graph as input context</td>
<td>ERNIE - contextualize entity emb</td>
</tr>
<tr>
<td></td>
<td>CoLAKE - contextualize KG triplet</td>
</tr>
<tr>
<td></td>
<td>DRAGON - contextualize KG subgraph</td>
</tr>
</tbody>
</table>
Takeaways

- Knowledge can be integrated into LM in self-supervised ways
- Help a wide range of reasoning tasks

Open questions

- Can we integrate knowledge in pre-finetuning (e.g. instruction tuning, RLHF)?
- How can we build a unified model with all various knowledge sources?
- How can we ensure the models use and reason about knowledge faithfully?
Thank you!

https://cs.stanford.edu/~myasu/

@michiyasunaga