CSCE 689 - Special Topics in NLP for Science

Lecture 17: Language Models with Academic Graphs

Yu Zhang

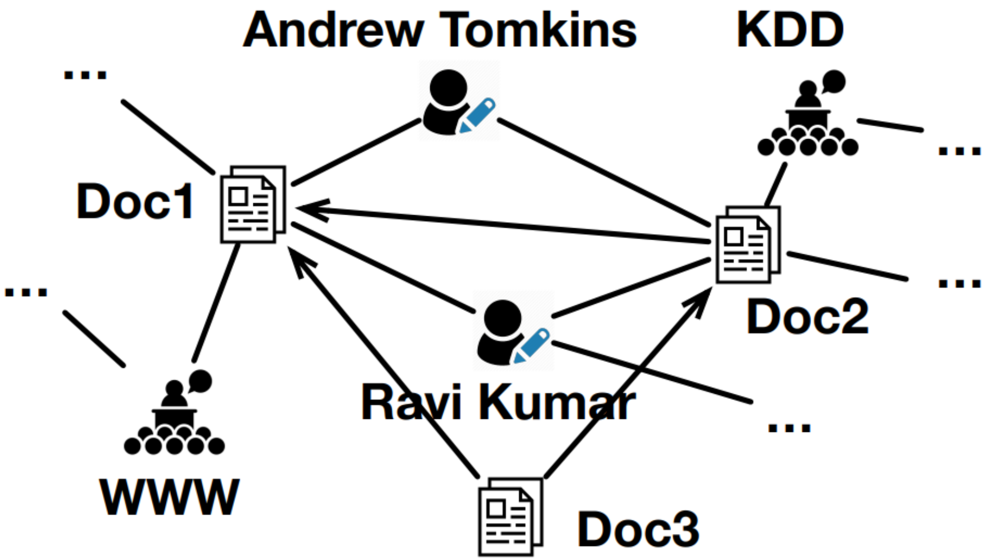
yuzhang@tamu.edu

March 20, 2025

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html

Academic Graphs

- Nodes: Paper, Venue, Author, (Year, Field, Affiliation, …)
- Edges: Paper→Paper, Paper-Venue, Paper-Author, (Author-Affiliation, …)
- Node Attributes: Paper title/abstract, Venue name, Author name, …



Are LLMs aware of academic graphs?

You

What is the most cited paper in WWW 2017? Please list its title and authors.

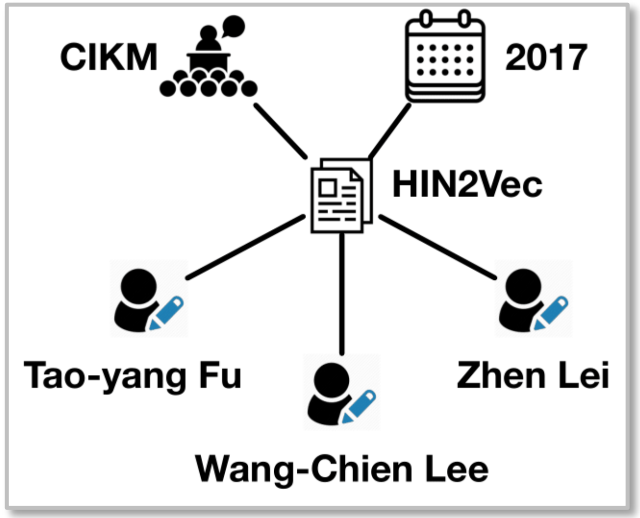
ChatGPT

As of my last update in January 2022, I don't have the real-time citation information, but one of the highly cited papers from WWW 2017 is:

Title: "HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning"

Authors: Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, Edward Y. Chang

Please verify the latest citation counts and updates from reliable sources or academic databases for the most current information.



HIN2Vec was published in CIKM 2017.

HIN2Vec was written by Tao-yang Fu, Wang-Chien Lee, and Zhen Lei.

ChatGPT 3.5, queried on January 23, 2024

Are LLMs aware of academic graphs?

What is the most cited paper in WWW 2017? Please list its title and authors.

As of January 2025, the most cited paper from the 26th International World Wide Web Conference (WWW 2017) is:

Title: "Attention Is All You Need"

Authors: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin

This seminal paper introduced the Transformer architecture, which has become foundational in the field of natural language processing and has significantly influenced subsequent research and applications.

ChatGPT 4o, queried on January 7, 2025

Transformer was published in NeurIPS 2017.

Are LLMs aware of academic graphs?

What is the most cited paper in WWW 2017? Please list its title and authors.

Identifying the most cited paper from the 2017 International World Wide Web Conference (WWW 2017) requires access to current citation data, which I cannot retrieve directly. However, based on available information, one of the highly influential papers from WWW 2017 is:

Title: "Inductive Representation Learning on Large Graphs"

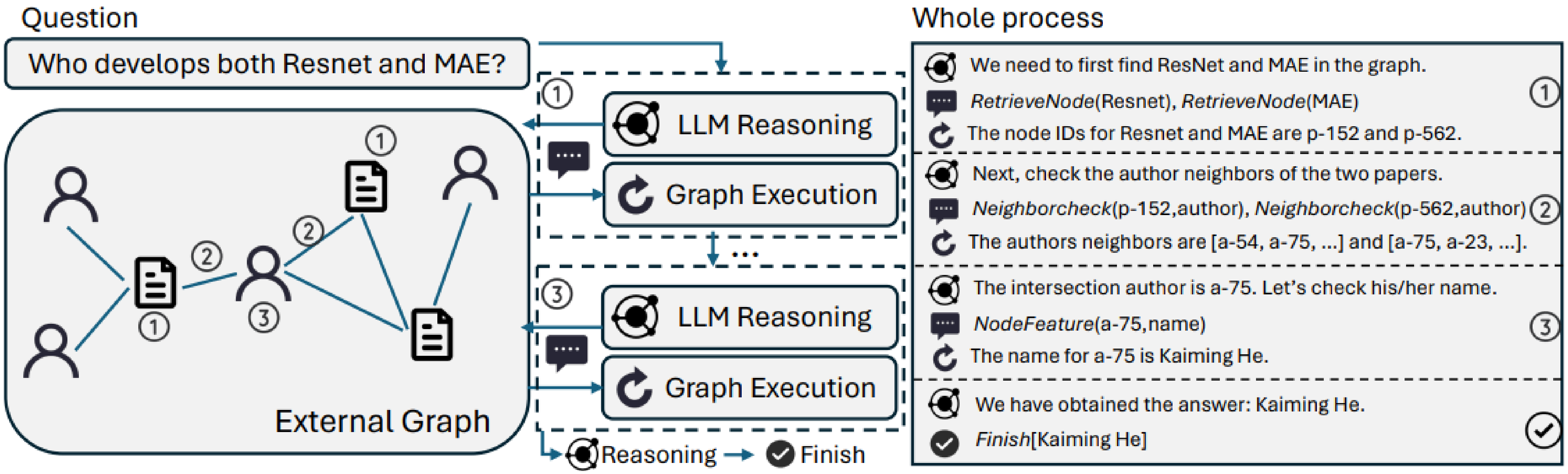
Authors: William L. Hamilton, Rex Ying, and Jure Leskovec

This paper introduces GraphSAGE, a framework for inductive representation learning on large graphs, which has been widely cited in subsequent research.

ChatGPT 4o, queried on March 15, 2025

GraphSAGE was published in NeurIPS 2017.

Last Lecture: Graph Chain-of-Thought



	Model	Academic		E-commerce		Literature		Healthcare		Legal	
		EM	GPT4score	EM	GPT4score	EM	GPT4score	EM	GPT4score	EM	GPT4score
Graph RAG	LLaMA-2-13b	22.01	22.97	12.48	20.00	9.25	20.00	2.97	4.81	17.98	17.22
Graph RAG	Mixtral-8x7b	27.77	31.20	32.87	37.00	20.08	33.33	8.66	15.19	23.48	25.56
Graph RAG	GPT-3.5-turbo	18.45	26.98	17.52	28.00	14.94	24.17	8.69	14.07	18.66	22.22
	GRAPH-COT	**31.89**	**33.48**	**42.40**	**44.50**	**41.59**	**46.25**	**22.33**	**28.89**	**30.52**	**28.33**

Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. ACL 2024 Findings.

Agenda

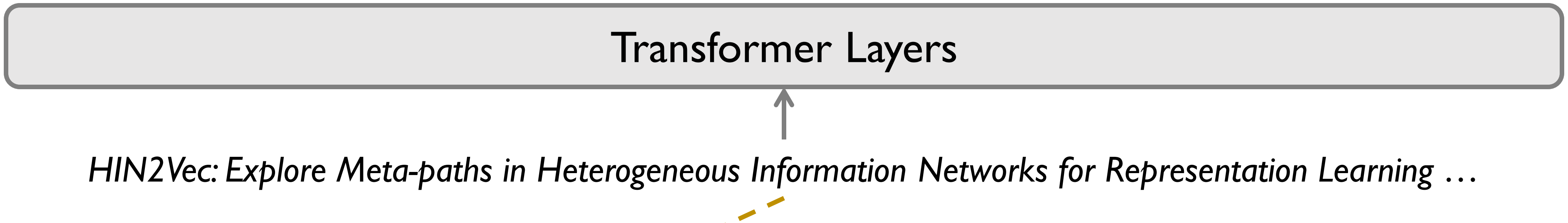
- Academic Graphs as Additional Input Features
  - OAG-BERT
  - LinkBERT
- Academic Graphs as Supervision
  - MICoL
  - GraphInst

Agenda

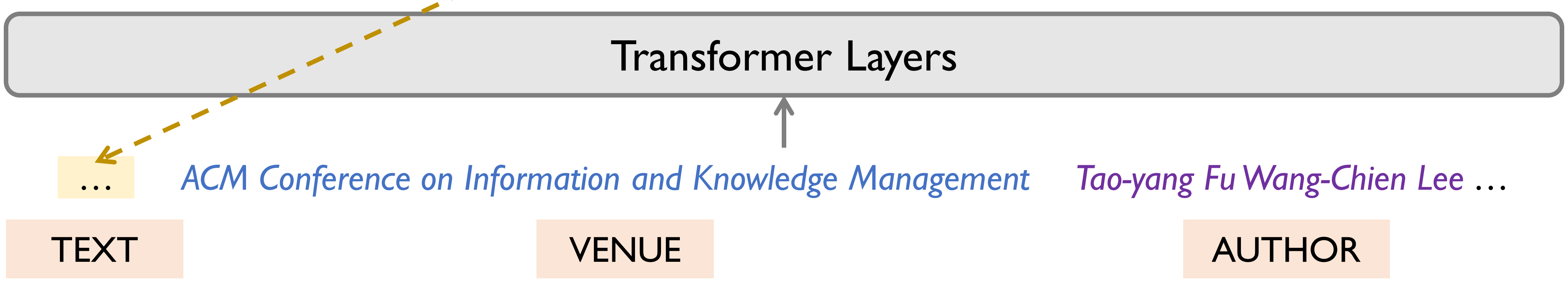
- Academic Graphs as Additional Input Features
  - OAG-BERT
  - LinkBERT
- Academic Graphs as Supervision
  - MICoL
  - GraphInst

How to sequentialize graph information?

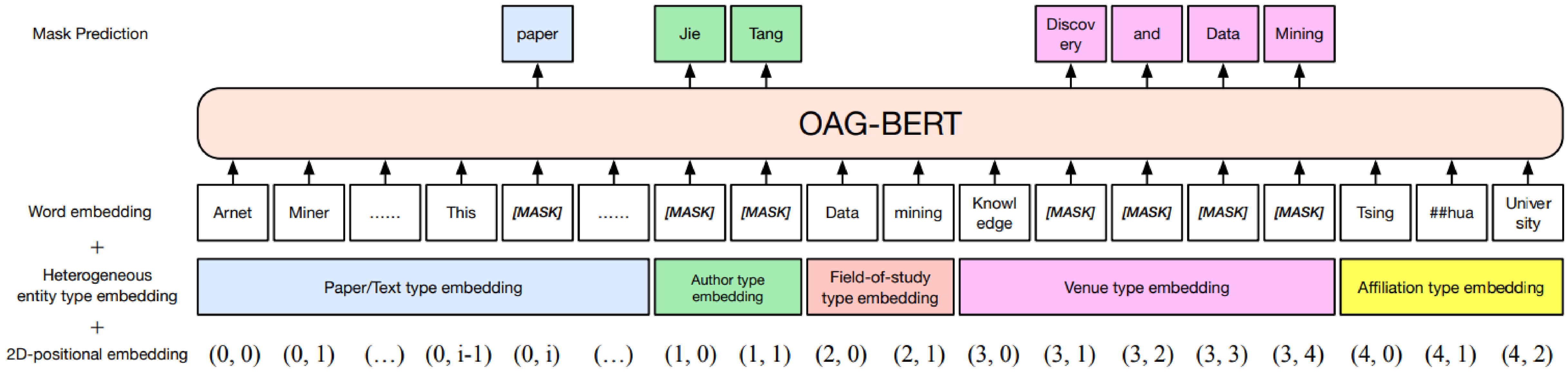
- BERT for Text



- BERT for Text + Graph



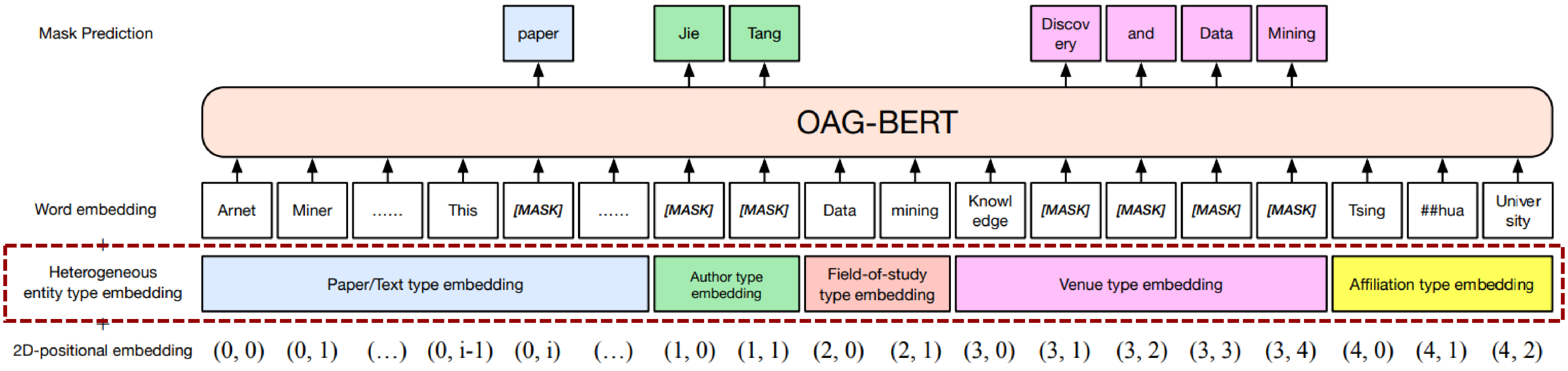
OAG-BERT: Overview



- Pre-train a new model for encoding an academic paper's text + graph information
- Append authors, fields-of-study, venue, and affiliation to paper text
- MLM only, no NSP

OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services. KDD 2022.

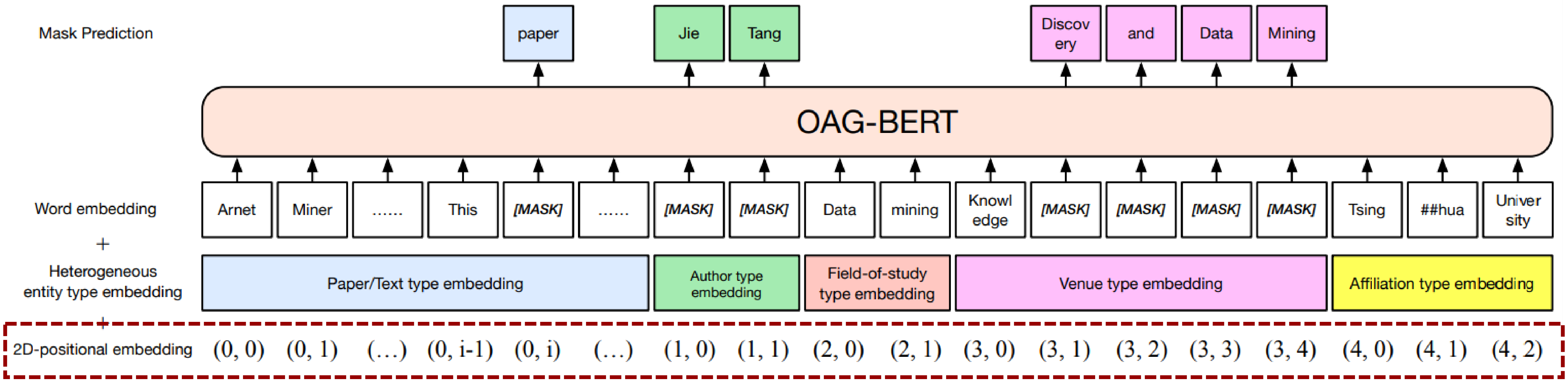
OAG-BERT: Heterogeneous Entity Type Embedding



- Make the model aware of different types of input
- Analogous to segment embeddings in the original BERT

OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services. KDD 2022.

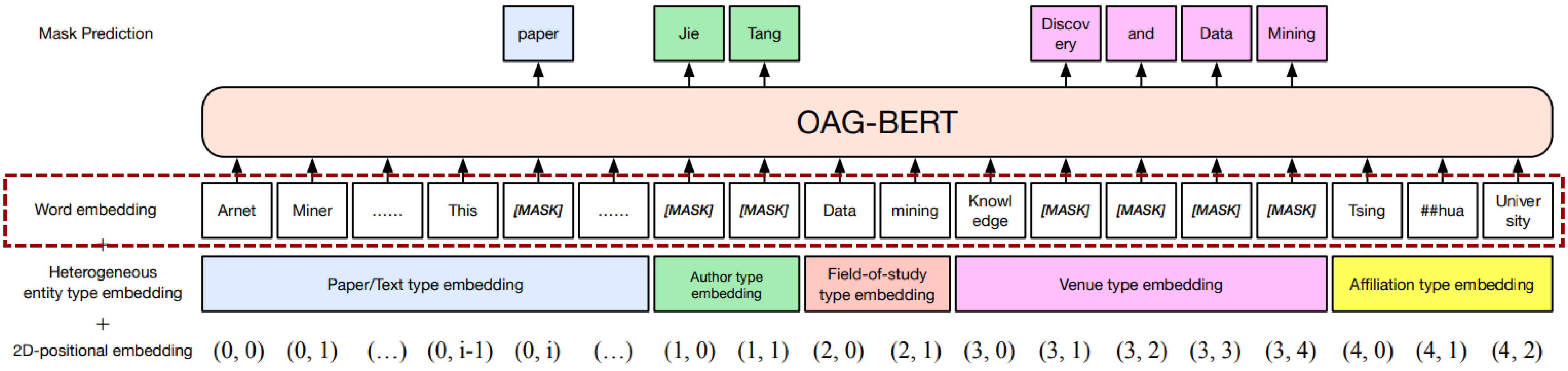
OAG-BERT: 2D-Positional Embedding



- Jointly model inter and intra-entity token orders
  - Further make aware of different types of input
- Final positional embedding = 1st positional embedding + 2nd positional embedding

OAG-BERT:Towards A Unified Backbone Language Model For Academic Knowledge Services. KDD 2022.

OAG-BERT: Span-Aware Entity Masking



- Select a continuous span when performing MLM on graph signals
    - <4 tokens: mask the entire entity
    - ≥ 4 tokens: randomly mask a continuous span of 4-10 tokens
- Motivation: In downstream applications, we usually need to predict the entire entity.

OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services. KDD 2022.

More Details of OAG-BERT

- Stage 1: MLM on text only
  - Data: AMiner + PubMed
- Stage 2: MLM on text + graph
  - Data: OAG (aligning AMiner and MAG)
- Averaging all token embeddings (instead of using [CLS]) as the document embedding

https://github.com/THUDM/OAG-BERT

README MIT license

```
from cogdl.oag import oagbert

tokenizer, bert_model = oagbert()

sequence = ["CogDL is developed by KEG, Tsinghua.", "OAGBert is developed by KEG, Tsinghua."]
tokens = tokenizer(sequence, return_tensors="pt", padding=True)
outputs = bert_model(**tokens)
```

Performance of OAG-BERT

- Author Name Disambiguation

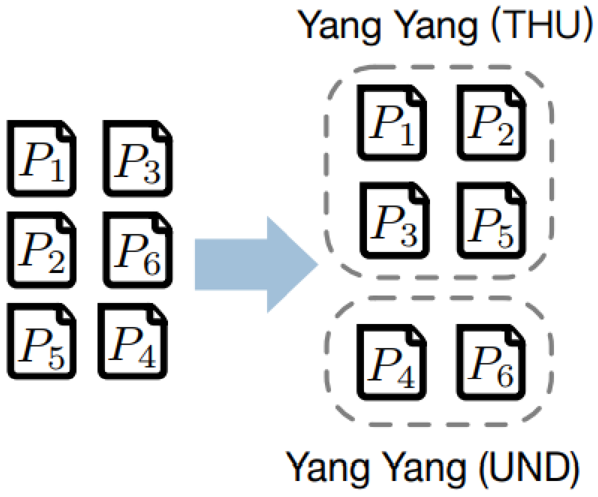


Table 1: The Macro Pairewise F1 scores for the author name disambiguation competition whoiswho-v1.

	Inputs	SciBERT	OAG-BERT
Unsupervised	*title*	0.3690	**0.4120**
	+fos	0.4101	**0.4643**
	+venue	0.3603	**0.4247**
	+fos+venue	0.3903	**0.4823**
Supervised	Leader Board Top1	0.4900	

- Literature Retrieval

Table 2: Scientific Literature Retrieval evaluation on OAG-QA (Top-100) between SciBERT and OAG-BERT.

	SciBERT	OAG-BERT
Geometry	0.097	**0.147**
Math. & Stats.	0.099	**0.166**
Algebra	**0.071**	0.069
Calculus	0.091	**0.160**
Number theory	0.067	**0.085**
Linear algebra	0.111	**0.160**
Astrophysics	0.041	**0.072**
Quantum mechanics	0.047	**0.080**
Classical mechanics	0.085	**0.197**
Chemistry	0.181	**0.216**
Biochemistry	0.146	**0.319**
Health care	0.041	**0.262**
Natural science	0.101	**0.277**
Algorithm	0.084	**0.209**
Neuroscience	0.054	**0.120**
Computer vision	0.035	**0.205**
Data mining	0.082	**0.161**
Deep learning	0.044	**0.138**
Machine learning	0.085	**0.177**
NLP	0.05	**0.160**
Economics	0.055	**0.151**
Average	0.079	**0.168**

Performance of OAG-BERT

- Link Prediction

Table 3: Paper recommendation and User Activity Prediction (Co-View and Co-Read) on Scidocs [8].

Models	Paper Rec.		Co-View		Co-Read	
	nDCG	P@1	MAP	nDCG	MAP	nDCG
Random	51.3	16.8	25.2	51.6	25.6	51.9
doc2vec	51.7	16.9	67.8	82.9	64.9	81.6
Sent-BERT	51.6	17.1	68.2	83.3	64.8	81.3
SciBERT	52.1	17.9	50.7	73.1	47.7	71.1
OAG-BERT	**52.6**	**18.6**	**74.7**	**86.3**	**71.4**	**84.7**

Still significantly underperforms SPECTER and SciNCL

- Paper Title Generation

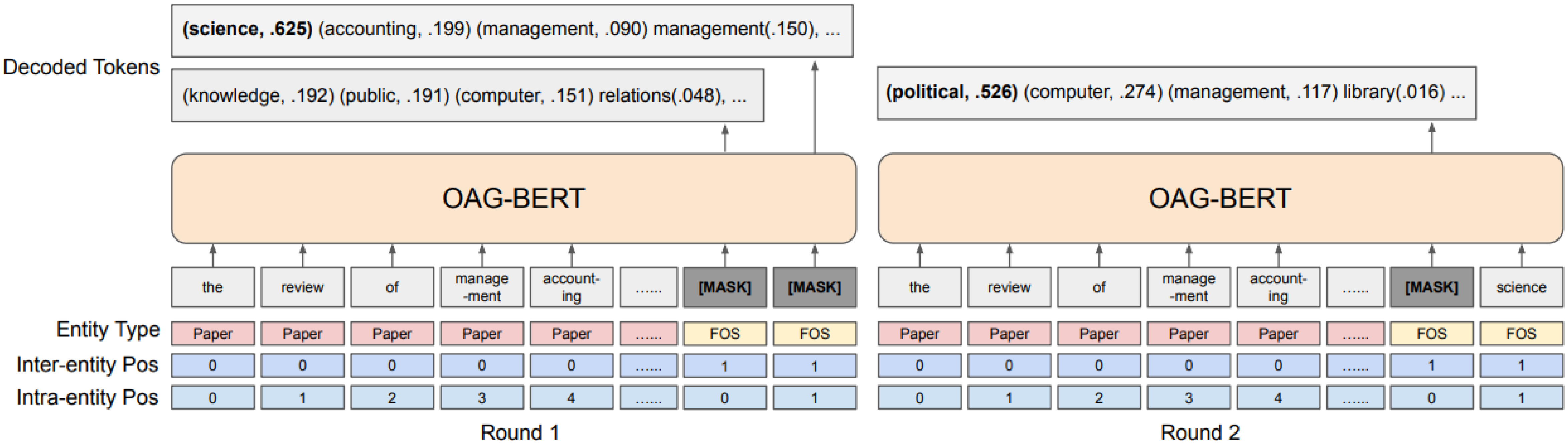
Table 7: Upper: case study in OAG-BERT generated titles and original title. **Lower:** Online testing result from 660 random human views on 50 pairs of OAG-BERT generated and original titles.

	OAG-BERT Generated v.s. Original
OAG-BERT	OAG-LM: A Unified Backbone for Academic Knowledge Services OAG-LM: A Unified Backbone Language Model for Academic Knowledge Services
AMiner	ArnetMiner: A System for Extracting and Mining Academic Social Networks ArnetMiner: Extraction and Mining of Academic Social Networks
ResNet	Deep Residual Networks for Visual Recognition : A Comparison of Deep and VGG Networks Deep Residual Networks for Image Recognition
SciBERT	SciBERT: A Pretrained Language Model for Scientific NLP SciBERT: A Pretrained Language Model for Scientific Text

Method	Total	Select	Selection Rate
OAG-BERT Generated	330	157	47.6%
Original	330	163	52.4%

Performance of OAG-BERT

- Metadata Prediction



- Enumerate the length (i.e., number of [MASK] tokens) and pick the answer with the highest conditional probability.

Performance of OAG-BERT

- Metadata Prediction

Table 5: The results for zero-shot inference tasks.

Method	Paper Tagging		Venue		Affiliation	
	Hit@1	MRR	Hit@1	MRR	Hit@1	MRR
SciBERT	19.93%	0.37	9.87%	0.22	6.93%	0.19
+prompt	29.59%	0.47	10.03%	0.21	8.00%	0.20
+abstract	25.66%	0.43	18.00%	0.32	10.33%	0.22
+both	35.33%	0.52	9.83%	0.22	12.40%	0.25
OAG-BERT	34.36%	0.51	21.00%	0.37	11.03%	0.24
+prompt	37.33%	0.55	22.67%	0.39	11.77%	0.25
+abstract	**49.59%**	**0.67**	**39.00%**	**0.57**	**21.67%**	**0.38**
+both	49.51%	**0.67**	38.47%	**0.57**	21.53%	**0.38**

Take-Away Messages

- Performing MLM jointly on text and graph signals (i.e., metadata neighbors) of a paper enhances the representation learning ability of the model.
    - The heterogeneous entity type embedding and the 2D-positional embedding make the model aware of different types of input.
- The model can predict metadata of a paper via zero-shot prompting.
    - But we need to enumerate the length of the metadata.

- Limitations:
    - Authors are intuitively semantic-indicative, but author names are practically hard to deal with.
        - Treat each author name as one token: How to deal with new authors? Explosion of the vocabulary size?
        - Tokenize the author names: Two authors sharing the same first/last name do not necessarily work on the same topic.

Agenda

- Academic Graphs as Additional Input Features
  - OAG-BERT
  - LinkBERT
- Academic Graphs as Supervision
  - MICoL
  - GraphInst

References/Citation Links

- Available in scientific papers, Wikipedia articles, webpages, …
- We have seen models using citation links as supervision (e.g., SPECTER and SciNCL).
- How to use them as additional features?
  - Not considered in OAG-BERT

An example where references benefit question answering

Document - - - - - - - - → **Linked document**

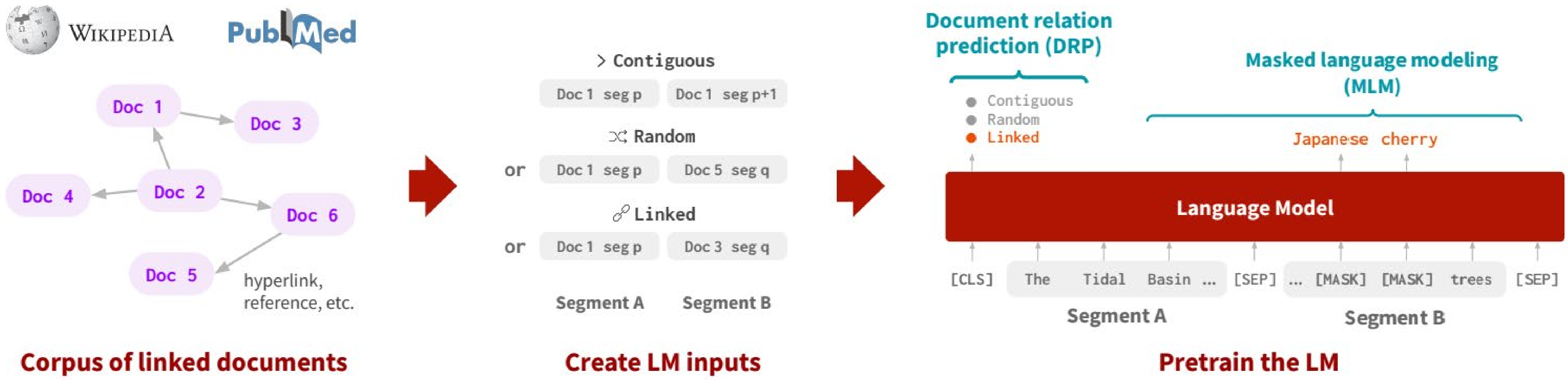
(*e.g.* hyperlink, reference)

[Tidal Basin, Washington D.C.]
The Tidal Basin is a man-made reservoir located between the Potomac River and the Washington Channel in Washington, D.C. 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 Martin Luther King Jr. Memorial, the Franklin Delano Roosevelt Memorial, and the George Mason Memorial are situated adjacent to the Tidal Basin.

[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 the growing friendship between the United States and Japan. ... Of the initial gift of 12 varieties of 3,020 trees, **the Yoshino Cherry** (70% of total) and Kwanzan Cherry (13% of total) now dominate. ...

LinkBERT: A Cross-Encoder Architecture

- BERT – A pair of sentences (next or random). Simultaneously perform MLM and NSP (binary classification).
- LinkBERT – A pair of sentences (next, random, or linked). Simultaneously perform MLM and NSP (three-class classification)

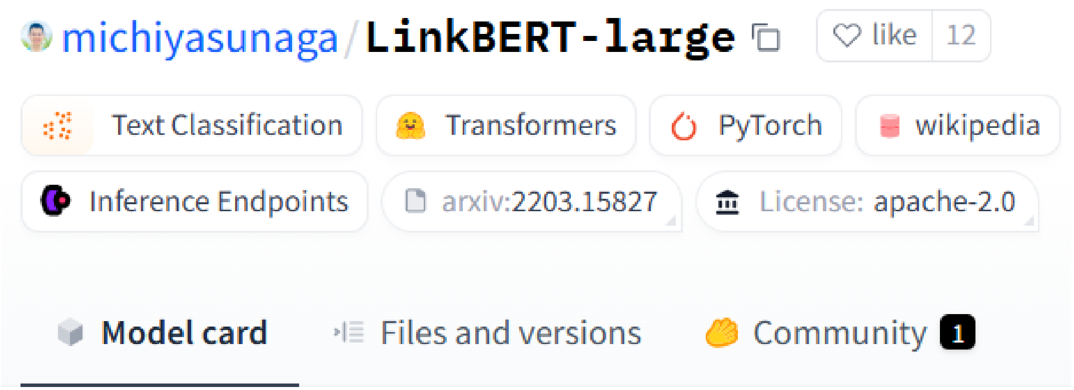


LinkBERT: Pretraining Language Models with Document Links. ACL 2022.

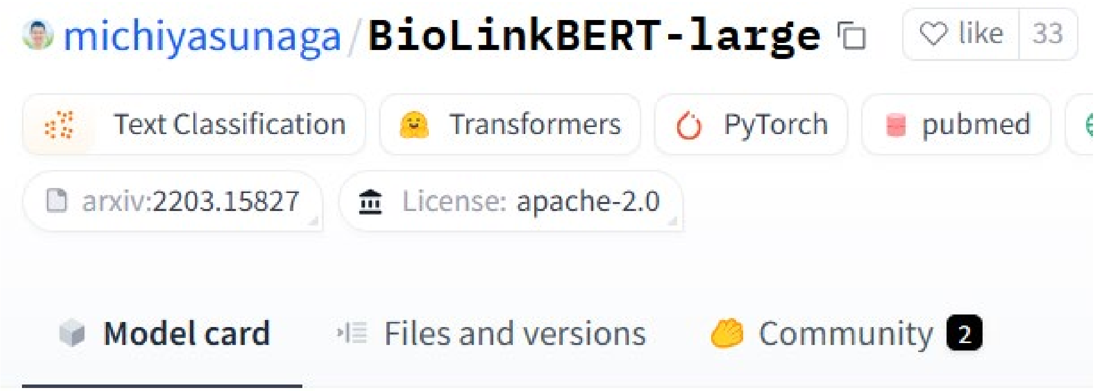
More Details of LinkBERT & BioLinkBERT

- LinkBERT: continue pre-training BERT-Tiny/Base/Large (4.4M/110M/340M parameters) using Wikipedia reference links
  - Consistently outperform BERT-Tiny/Base/Large on various extractive QA datasets
- BioLinkBERT: pre-training a base/large-size BERT model (110M/340M parameters) from scratch using PubMed reference links

https://huggingface.co/michiyasunaga/LinkBERT-large



https://huggingface.co/michiyasunaga/BioLinkBERT-large



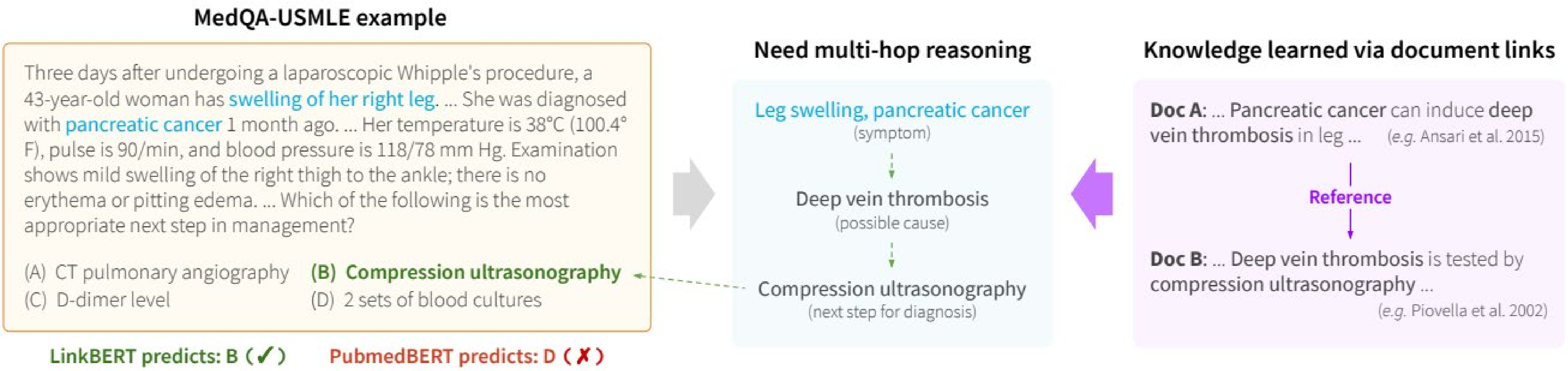
Performance of BioLinkBERT: Multi-Choice QA

MedQA-USMLE

Methods	Acc. (%)
BioBERT$_{\texttt{large}}$ (Lee et al., 2020)	36.7
QAGNN (Yasunaga et al., 2021)	38.0
GreaseLM (Zhang et al., 2022)	38.5
PubmedBERT$_{\texttt{base}}$ (Gu et al., 2020)	38.1
BioLinkBERT$_{\texttt{base}}$ (**Ours**)	**40.0**
BioLinkBERT$_{\texttt{large}}$ (**Ours**)	**44.6**

MMLU-Professional-Medicine

Methods	Acc. (%)
GPT-3 (175B params) (Brown et al., 2020)	38.7
UnifiedQA (11B params) (Khashabi et al., 2020)	43.2
BioLinkBERT$_{\texttt{large}}$ (**Ours**)	**50.7**



Performance of BioLinkBERT: BLURB Benchmark

	PubMed-BERT$_{\text{base}}$	BioLink-BERT$_{\text{base}}$	BioLink-BERT$_{\text{large}}$
Named entity recognition			
BC5-chem (Li et al., 2016)	93.33	**93.75**	**94.04**
BC5-disease (Li et al., 2016)	85.62	**86.10**	**86.39**
NCBI-disease (Doğan et al., 2014)	87.82	**88.18**	**88.76**
BC2GM (Smith et al., 2008)	84.52	**84.90**	**85.18**
JNLPBA (Kim et al., 2004)	**80.06**	79.03	**80.06**
PICO extraction			
EBM PICO (Nye et al., 2018)	73.38	**73.97**	**74.19**
Relation extraction			
ChemProt (Krallinger et al., 2017)	77.24	**77.57**	**79.98**
DDI (Herrero-Zazo et al., 2013)	82.36	**82.72**	**83.35**
GAD (Bravo et al., 2015)	82.34	**84.39**	**84.90**
Sentence similarity			
BIOSSES (Soğancıoğlu et al., 2017)	92.30	**93.25**	**93.63**
Document classification			
HoC (Baker et al., 2016)	82.32	**84.35**	**84.87**
Question answering			
PubMedQA (Jin et al., 2019)	55.84	**70.20**	**72.18**
BioASQ (Nentidis et al., 2019)	87.56	**91.43**	**94.82**
BLURB score	81.10	**83.39**	**84.30**

https://microsoft.github.io/BLURB/leaderboard.html

BLURB

The Overall score is calculated as the macro-average performance over tas

Show 100 entries

Rank	Model	BLURB Score (Macro Avg.)
1	**BioLinkBERT-Large** — Stanford	**84.30**
2	**BioM-ALBERT-xxlarge-PMC** — University of Delaware	84.10
3	**BioM-ELECTRA-Large** — University of Delaware	83.81

Take-Away Messages

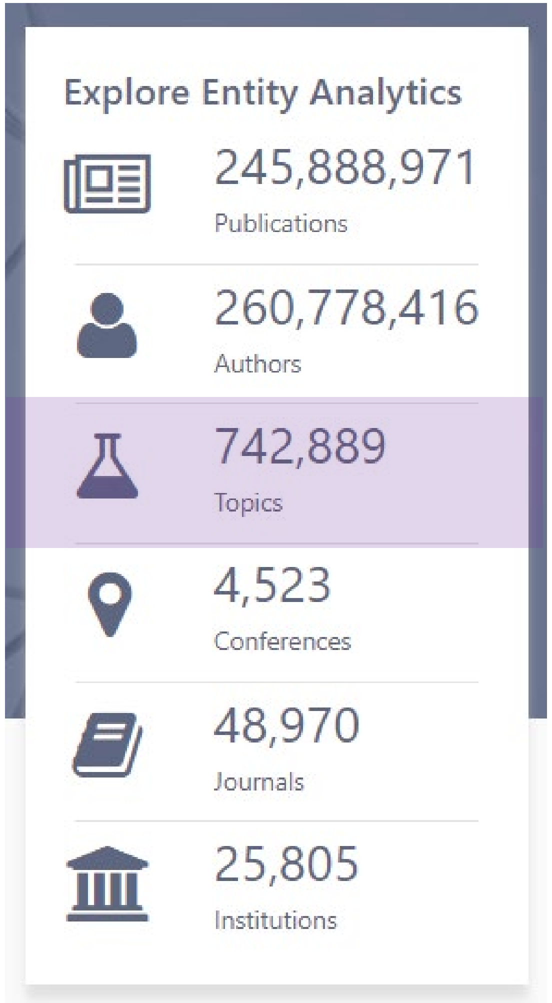
- Adding a sentence from a reference paper as context helps capture knowledge and semantics not reflected in the local context within each paper.
- Such a cross-encoder pre-training paradigm consistently benefits QA tasks.
  - Why?
  - Extractive QA – Input: Context [SEP] Question; Goal: Find information (i.e., a span of tokens) in the context that is relevant to the question
  - LinkBERT pre-training – Input: Paper 1 [SEP] Paper 2; Goal: Judge if there is information in Paper 1 that is relevant to Paper 2 (which may imply that Paper 1 cites Paper 2)

- Limitation:
  - LinkBERT pre-training models only one citation edge every time. Is it possible to include all references simultaneously as additional features?
    - *MATCH: Metadata-Aware Text Classification in a Large Hierarchy.* WWW 2021.
  - How to further include other graph signals (e.g., author, venue, etc.)?

Agenda

- Academic Graphs as Additional Input Features
  - OAG-BERT
  - LinkBERT
- Academic Graphs as Supervision
  - MICoL
  - GraphInst

Extremely Fine-Grained Scientific Paper Classification



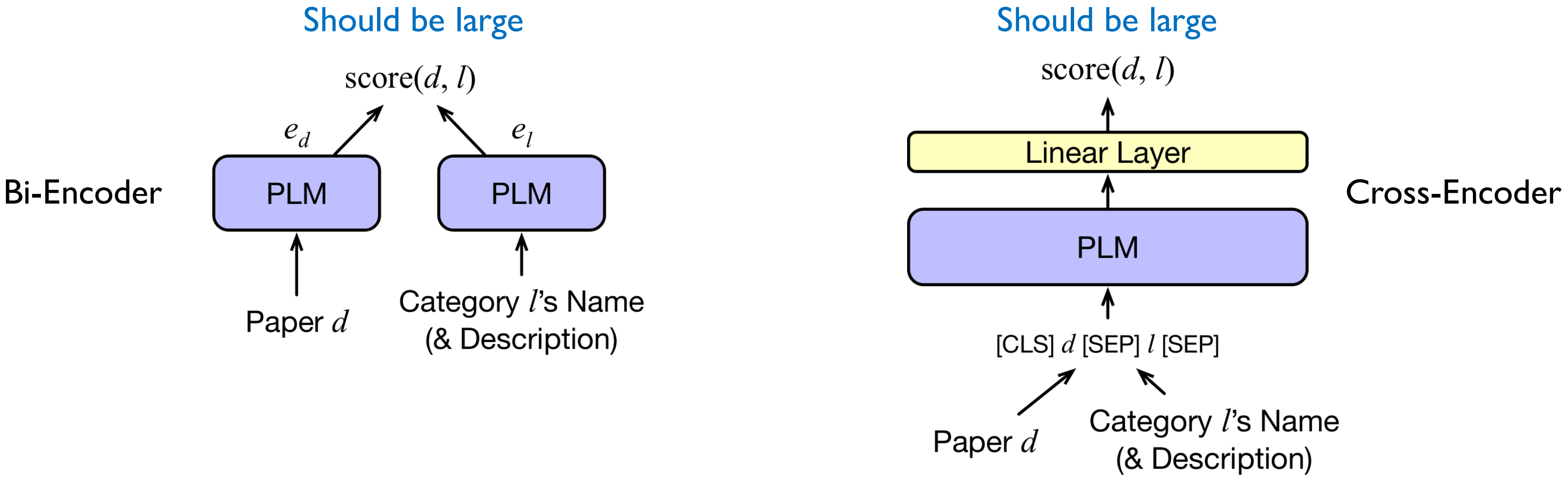
- MAG has 740K+ categories.
- The Medical Subject Headings (MeSH) for indexing PubMed papers contain 30K+ categories.
- Each paper can be relevant to more than one category (5-15 categories for most papers).

Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.

- Relevant categories: Betacoronavirus, Cardiovascular Diseases, Comorbidity, Coronavirus Infections, Fibrin Fibrinogen Degradation Products, Mortality, Pandemics, Patient Isolation, Pneumonia, …

If we could have some training data …

- We could use relevant (paper, category) pairs to fine-tune a pre-trained language model.
- Both Bi-Encoder and Cross-Encoder are applicable.

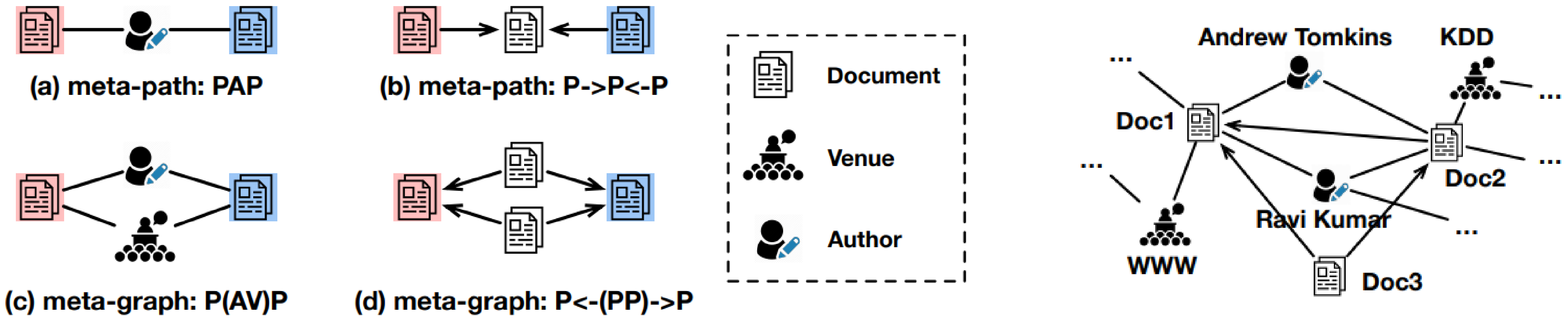


- However, human-annotated training samples are NOT available in many cases!
  - We are asking annotators to find ~10 relevant categories from ~100,000 candidates!

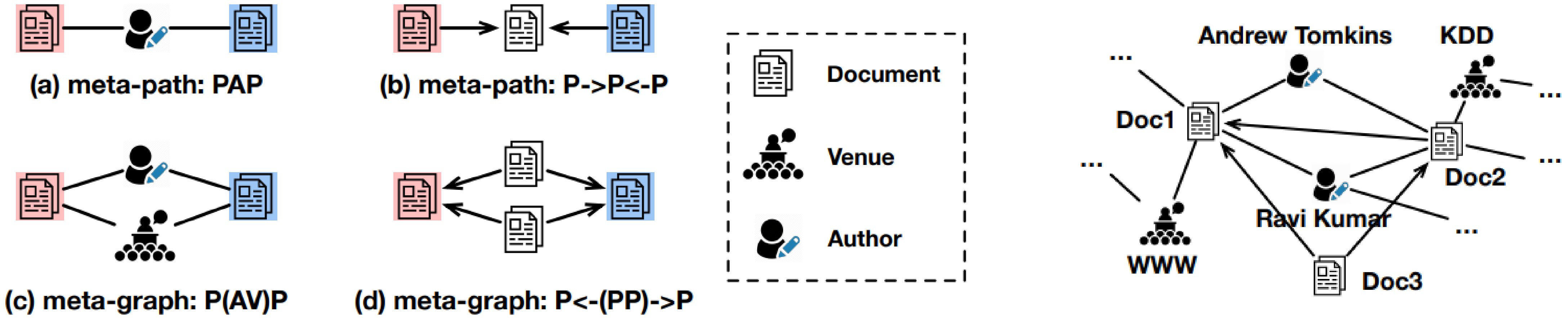
Using Academic Graph Signals to Replace Annotations

- If relevant (paper, category) pairs are not available, can we automatically create relevant (paper, paper) pairs?
  - Two papers sharing the same author(s) are assumed to be similar.
  - Two papers sharing the same reference(s) are assumed to be similar.
  - …

- The notion of meta-paths and meta-graphs



Using Academic Graph Signals to Replace Annotations

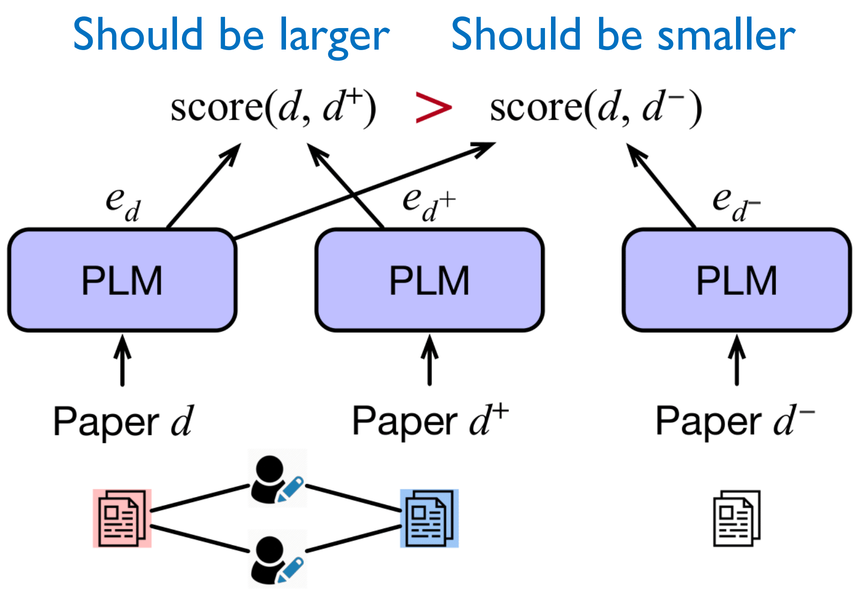


- Examples:
  - Doc1 and Doc2 are connected via the meta-path PAP.
  - Doc1 and Doc2 are NOT connected via the meta-path PVP.
  - Doc1 and Doc2 are connected via the meta-graph P(AA)P.
- Why do need to consider meta-graphs?
  - One author may work on many different topics, but when two authors collaborate, the scope usually becomes much narrower.

Graph-Induced Contrastive Learning

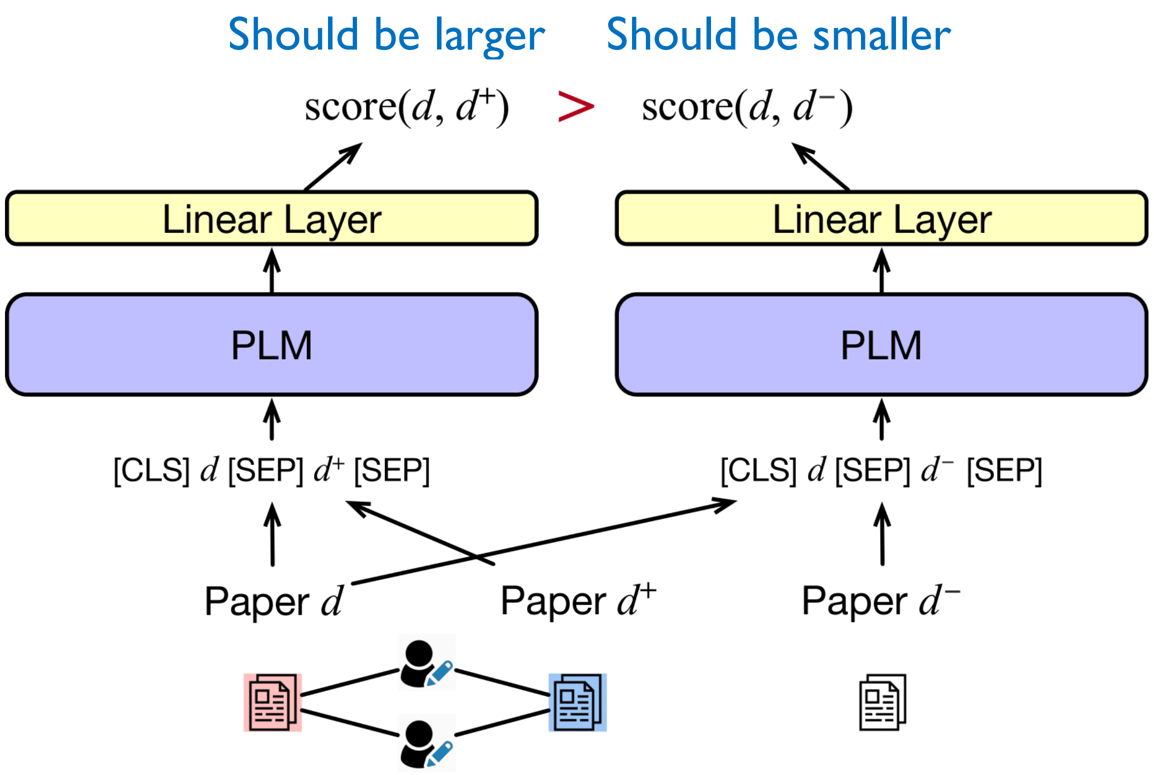
- Two papers connected via a certain meta-path/meta-graph should be more similar than two randomly selected papers.

Bi-Encoder



$$-\log \frac{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau)}{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau) + \sum_{i=1}^{N} \exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d_i^-})/\tau)}$$

Cross-Encoder



Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification. WWW 2022.

Performance of MICoL

Table 2: P@k and NDCG@k scores of compared algorithms on MAG-CS and PubMed. Bold: the highest score of zero-shot approaches. *: MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$) is significantly better than this algorithm with p-value < 0.05. **: MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$) is significantly better than this algorithm with p-value < 0.01.

	Algorithm	MAG-CS [49]					PubMed [24]				
		P@1	P@3	P@5	NDCG@3	NDCG@5	P@1	P@3	P@5	NDCG@3	NDCG@5
Zero-shot	Doc2Vec [31]	0.5697**	0.4613**	0.3814**	0.5043**	0.4719**	0.3888**	0.3283**	0.2859**	0.3463**	0.3252**
	SciBERT [2]	0.6440**	0.5030**	0.4011**	0.5545**	0.5061**	0.4427**	0.3572**	0.3031**	0.3809**	0.3510**
	ZeroShot-Entail [61]	0.6649**	0.5003**	0.3959**	0.5570**	0.5057**	0.5275**	0.4021	**0.3299**	0.4352	**0.3913**
	SPECTER [8]	0.7107**	0.5381**	0.4184**	0.5979**	0.5365**	0.5286**	0.3923**	0.3181**	0.4273**	0.3815**
	EDA [53]	0.6442**	0.4939**	0.3948**	0.5471**	0.5000**	0.4919	0.3754*	0.3101*	0.4058*	0.3667*
	UDA [57]	0.6291**	0.4848**	0.3897**	0.5362**	0.4918**	0.4795**	0.3696**	0.3067**	0.3986**	0.3614**
	MICoL (Bi-Encoder, $P \rightarrow P \leftarrow P$)	0.7062*	0.5369*	0.4184*	0.5960*	0.5355*	0.5124**	0.3869*	0.3172*	0.4196*	0.3774*
	MICoL (Bi-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.7050*	0.5344*	0.4161*	0.5937*	0.5331*	0.5198**	0.3876*	0.3172*	0.4215*	0.3786*
	MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$)	**0.7177**	**0.5444**	**0.4219**	**0.6048**	**0.5415**	**0.5412**	**0.4036**	0.3257	**0.4391**	0.3906
	MICoL (Cross-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.7061	0.5376	0.4187	0.5964	0.5357	0.5218	0.3911	0.3172*	0.4249	0.3794

- Cross-Encoder > Bi-Encoder
- Meta-paths > One-hop citation information (SPECTER)
- Graph-induced contrastive learning > Text-only contrastive learning (EDA & UDA)

Effect of Meta-Paths/Meta-Graphs

Algorithm	MAG-CS [49]					PubMed [24]				
	P@1	P@3	P@5	NDCG@3	NDCG@5	P@1	P@3	P@5	NDCG@3	NDCG@5
Unfine-tuned SciBERT	0.6599**	0.5117**	0.4056**	0.5651**	0.5136**	0.4371**	0.3544**	0.3014**	0.3775**	0.3485**
MICoL (Bi-Encoder, PAP)	0.6877**	0.5285**	0.4143**	0.5852**	0.5280**	0.4974**	0.3818**	0.3154*	0.4122**	0.3727**
MICoL (Bi-Encoder, PVP)	0.6589**	0.5123**	0.4063**	0.5656**	0.5145**	0.4440**	0.3507**	0.2966**	0.3761**	0.3458**
MICoL (Bi-Encoder, $P \rightarrow P$)	0.7094	0.5391	0.4190	0.5982	0.5367	0.5200*	0.3903*	0.3195	0.4240*	0.3808*
MICoL (Bi-Encoder, $P \leftarrow P$)	0.7095*	0.5374*	0.4178*	0.5970*	0.5356*	0.5195**	0.3905*	0.3192	0.4240*	0.3806*
MICoL (Bi-Encoder, $P \rightarrow P \leftarrow P$)	0.7062*	0.5369*	0.4184*	0.5960*	0.5355*	0.5124**	0.3869*	0.3172*	0.4196*	0.3774*
MICoL (Bi-Encoder, $P \leftarrow P \rightarrow P$)	0.7039*	0.5379*	0.4187*	0.5963*	0.5356*	0.5174**	0.3886*	0.3187*	0.4220*	0.3795*
MICoL (Bi-Encoder, $P(AA)P$)	0.6873**	0.5272**	0.4130**	0.5840**	0.5269**	0.4963**	0.3794**	0.3139**	0.4101**	0.3711**
MICoL (Bi-Encoder, $P(AV)P$)	0.6832**	0.5263**	0.4135**	0.5823**	0.5263**	0.4894**	0.3743**	0.3099**	0.4045**	0.3664**
MICoL (Bi-Encoder, $P \rightarrow (PP) \leftarrow P$)	0.7015**	0.5334**	0.4160**	0.5920**	0.5322**	0.5163**	0.3879*	0.3172*	0.4211*	0.3781*
MICoL (Bi-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.7050*	0.5344*	0.4161*	0.5937*	0.5331*	0.5198**	0.3876*	0.3172*	0.4215*	0.3786*
MICoL (Cross-Encoder, PAP)	0.7034*	0.5355	0.4168	0.5943	0.5337	0.5212**	0.3921*	0.3207	0.4255*	0.3818*
MICoL (Cross-Encoder, PVP)	0.6720*	0.5203*	0.4103*	0.5750*	0.5210*	0.4668**	0.3633**	0.3051**	0.3908**	0.3574**
MICoL (Cross-Encoder, $P \rightarrow P$)	0.7033*	0.5391	0.4201	0.5971*	0.5365*	0.5266	0.3946	0.3207	0.4286	0.3830
MICoL (Cross-Encoder, $P \leftarrow P$)	0.7169	0.5430	0.4214	0.6033	0.5406	0.5265	0.3924	0.3186	0.4268	0.3811
MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$)	**0.7177**	**0.5444**	**0.4219**	**0.6048**	**0.5415**	**0.5412**	**0.4036**	**0.3257**	**0.4391**	**0.3906**
MICoL (Cross-Encoder, $P \leftarrow P \rightarrow P$)	0.7045	0.5356*	0.4168*	0.5944*	0.5336*	0.5243*	0.3932*	0.3190*	0.4271*	0.3814*
MICoL (Cross-Encoder, $P(AA)P$)	0.7028	0.5351	0.4171	0.5939	0.5338	0.5290*	0.3937	0.3201	0.4285*	0.3830
MICoL (Cross-Encoder, $P(AV)P$)	0.7024*	0.5354*	0.4177	0.5940*	0.5343*	0.5164**	0.3897*	0.3195*	0.4225*	0.3797*
MICoL (Cross-Encoder, $P \rightarrow (PP) \leftarrow P$)	0.7076*	0.5379*	0.4188	0.5971*	0.5363*	0.5186	0.3924*	0.3184*	0.4254*	0.3800*
MICoL (Cross-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.7061	0.5376	0.4187	0.5964	0.5357	0.5218	0.3911	0.3172*	0.4249	0.3794

- All examined meta-paths/meta-graphs are beneficial, except PVP.

Performance on Tail Labels

- New evaluation metrics: Propensity-Scored P/NDCG@k (a.k.a., PSP@k and PSN@k)
- If you can predict an infrequent label (e.g., "Lagrangian SVM") correctly, you will get a higher "reward" than you predict a frequent label (e.g., "Computer Science").

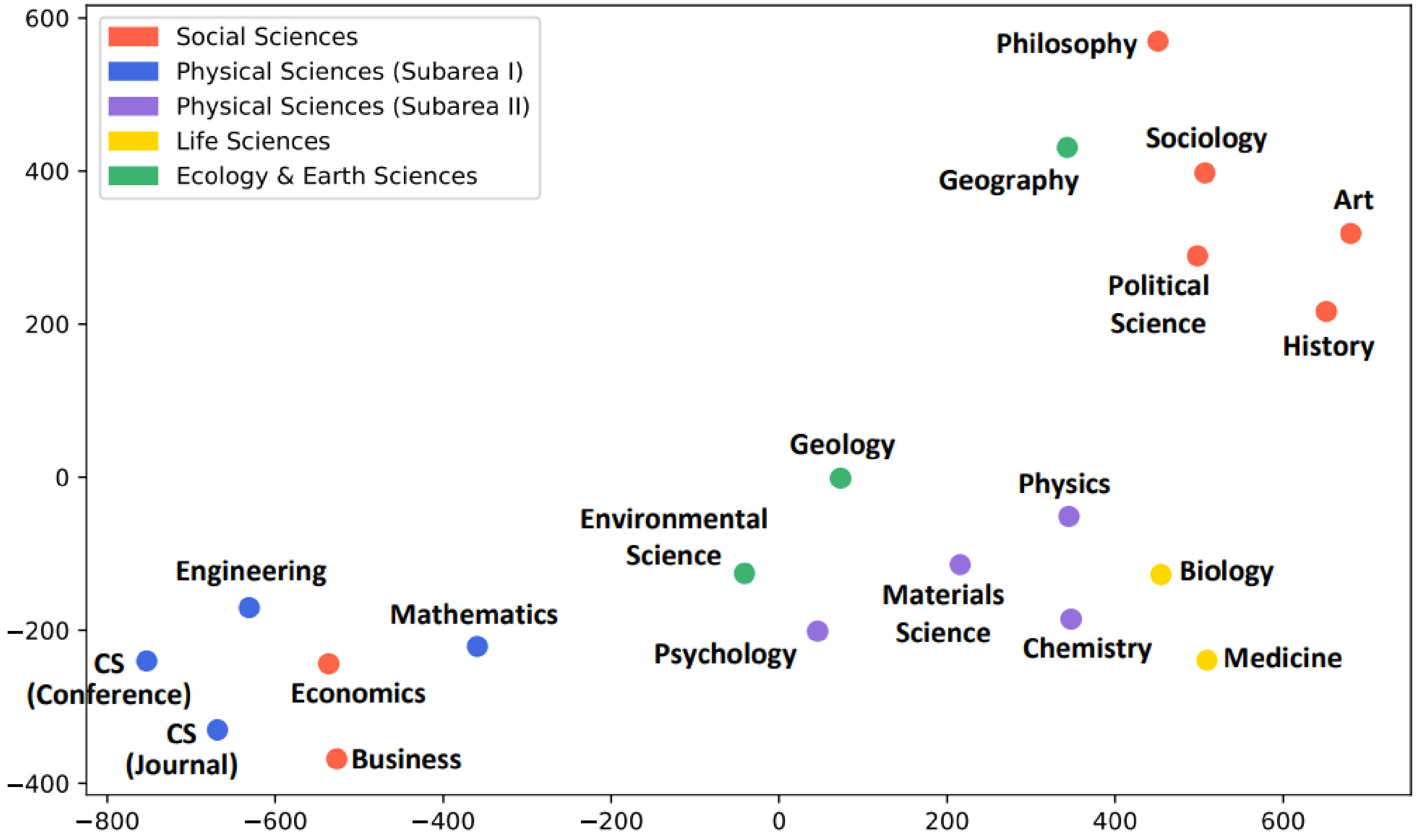
	Algorithm	MAG-CS [49]						PubMed [24]					
		PSP@1	PSP@3	PSP@5	PSN@3	PSN@5	PSP@1 / P@1	PSP@1	PSP@3	PSP@5	PSN@3	PSN@5	PSP@1 / P@1
Zero-shot	Doc2Vec [31]	0.4287**	0.4623**	0.4656**	0.4450**	0.4425**	0.75	0.2717**	0.2948**	0.3029**	0.2856**	0.2879**	0.70
	SciBERT [2]	0.4668**	0.4958**	0.4843**	0.4788**	0.4667**	0.72	0.3149**	0.3231**	0.3221**	0.3174**	0.3131**	0.71
	ZeroShot-Entail [61]	0.4796**	0.4892**	0.4759**	0.4777**	0.4644**	0.72	0.3617**	0.3498**	0.3389**	0.3492**	0.3378**	0.69
	SPECTER [8]	0.5304	0.5334*	0.5059*	0.5223	0.4988*	0.75	0.3907**	0.3638**	0.3442**	0.3666**	0.3489**	0.74
	EDA [53]	0.4916**	0.4968**	0.4821**	0.4859**	0.4708**	0.76	0.3572*	0.3451*	0.3334*	0.3442*	0.3322*	0.73
	UDA [57]	0.4850**	0.4907**	0.4771**	0.4797**	0.4654**	0.77	0.3547**	0.3423**	0.3311**	0.3416**	0.3298**	0.74
	MICoL (Bi-Encoder, $P \rightarrow P \leftarrow P$)	0.5176	0.5311	0.5065	0.5175	0.4963	0.73	0.3676**	0.3559**	0.3423*	0.3550**	0.3418**	0.72
	MICoL (Bi-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.5160	0.5281	0.5037	0.5150	0.4940	0.73	0.3780**	0.3589*	0.3423*	0.3597**	0.3450**	0.73
	MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$)	**0.5375**	**0.5415**	**0.5118**	**0.5302**	**0.5052**	0.75	**0.4105**	**0.3807**	**0.3558**	**0.3841**	**0.3625**	0.76
	MICoL (Cross-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.5326	0.5363	0.5087	0.5249	0.5013	0.75	0.3871	0.3664	0.3462	0.3677	0.3496	0.74

Take-Away Messages

- Given an academic graph, we can go beyond one-hop citation links to create positive (paper, paper) pairs for contrastive learning.
- For the fine-grained paper classification task, using venue information as supervision is too vague.
  - However, using venue as additional features is consistently helpful!
  - *The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study.* WWW 2023.
- Limitations
  - Rely on human knowledge to select a good meta-path/meta-graph. How do we know which meta-path/meta-graph is the most helpful?

Which type of metadata is the most helpful?

- Is the contribution of venues, authors, and references to paper classification consistent across different fields?
  - NO! BUT the effects of metadata tend to be similar in two similar fields.
  - The experience of using metadata in one field can be extrapolated to a similar field.



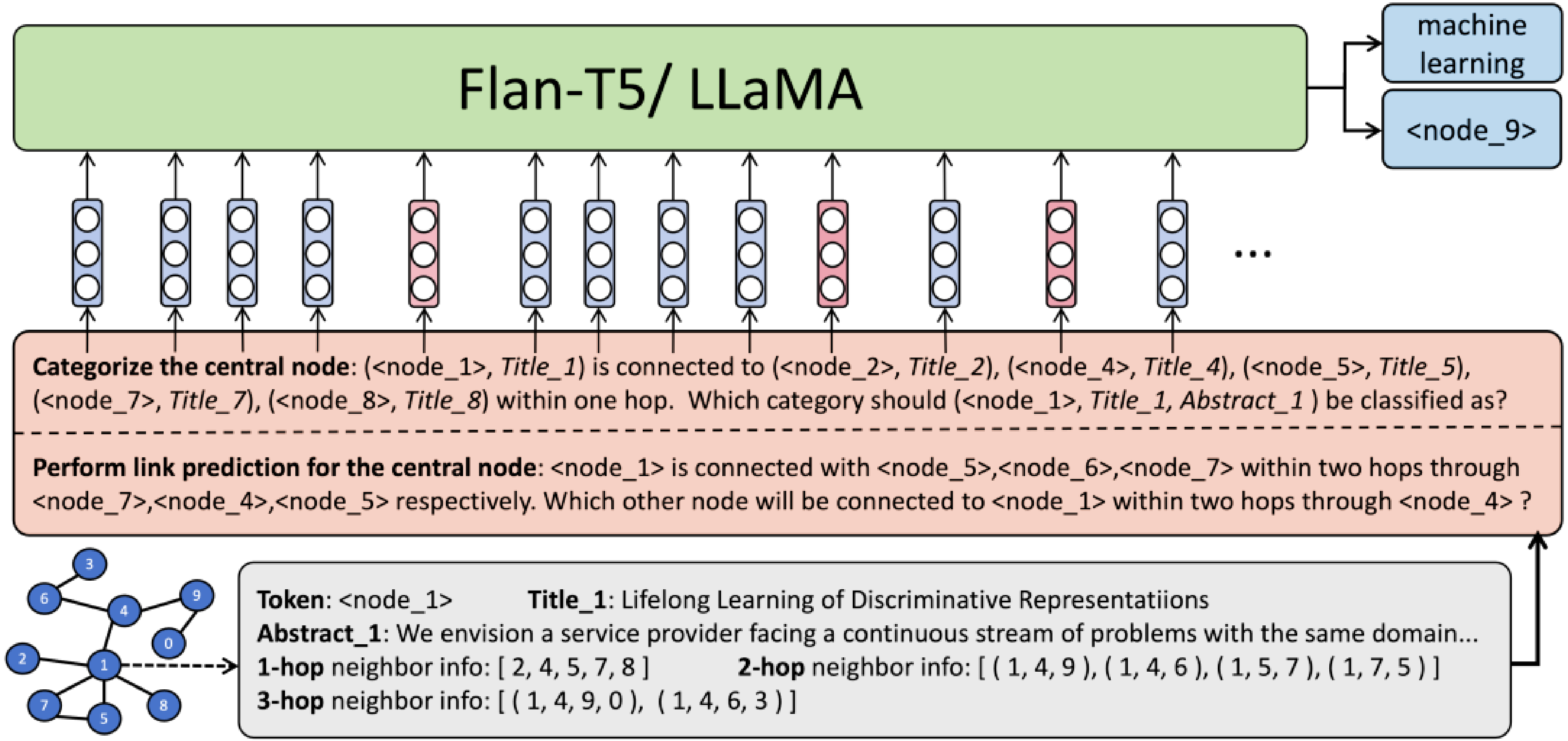
The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study. WWW 2023.

Agenda

- Academic Graphs as Additional Input Features
  - OAG-BERT
  - LinkBERT
- Academic Graphs as Supervision
  - MICoL
  - GraphInst

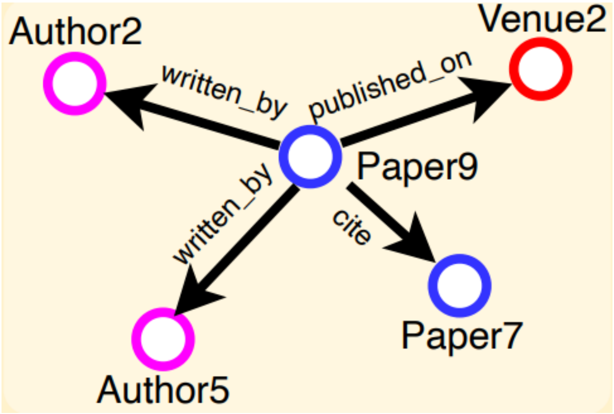
Instruction Tuning LLMs on Graphs

- How to teach an LLM about an unseen graph?



Language is All a Graph Needs. EACL 2024 Findings.

How to represent a graph?



- Natural Language
  - Paper9 is written by Author2 and Author5. Paper9 is published on Venue 2. Paper9 cites Paper 7 and …
- JSON
  - {“Paper9": {“written_by": [“Author2", “Author5"], “published_on": “Venue2”, “cite”: [“Paper7”, …]}}
- DOT (Code)
  - digraph G {

    Paper9 -> {Author2, Author5} [label=“written_by"];

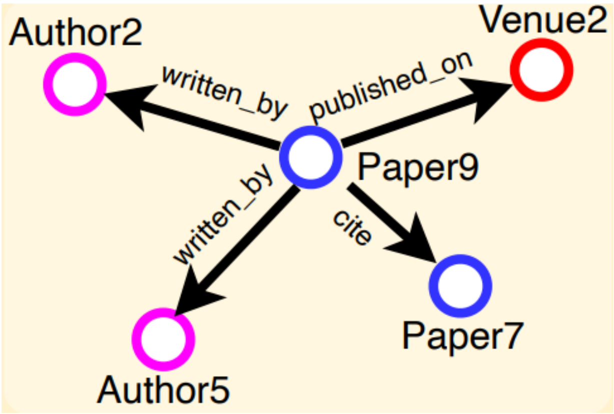
    Paper9 -> {Venue2} [label=“published_on"];

    Paper9 -> {Paper7, …} [label=“cite"]

  }

Investigating Instruction Tuning Large Language Models on Graphs. COLM 2024.

How should the LLM generalize?



- Seen Question
  - Q: What are the written_by neighbors of Paper9?
  - A: Author2, Author5
- Unseen Subtask
  - Q: What is the published_on neighbor of Paper9?
  - A: Venue2
- Unseen Answer Type
  - Q: How many written_by neighbors does Paper9 have?
  - A: 2
- Unseen Domain (Given a new graph, e.g., medical KG)
  - Q: What are the caused_by neighbors of Disease25?
  - A: Chemical6, Chemical10

Investigating Instruction Tuning Large Language Models on Graphs. COLM 2024.

Examined Answer Types, Tasks, and Subtasks

Answer Type	Task	Description
Node	Find neighbors	$\{v \in N(u) \mid \phi_E(u,v) \in T'_E\}$
	Nodes shared neighbors	$\{v \mid \forall t_e \in T'_E, \exists w \in V, \phi_E(u,w) = \phi_E(v,w) = t_e\}$
	N-hop neighbors	$\{v \mid \phi_V(v) \in T'_V, d(u,v) <= c\}$
Pair	Find pairs	$\{(v_1,v_2) \mid \phi(v_1,v_2) \in T'_E, u \in \{v_1,v_2\}\}$
	Pairs shared neighbors	$\{(v_1,v_2) \mid \exists W \subset V : \forall w \in W, \phi_E(v_1,w) = \phi_E(v_2,w) \in T'_E \wedge \lvert W \rvert = c\}$
Count	Degree count	$\{\lvert V' \rvert \mid V' \subseteq N(u) : \forall v \in V', \phi_E(u,v) \in T'_E\}$
	Node count within N-hop	$\{\lvert V' \rvert \mid V' \subset V : \forall v \in V', d(u,v) <= c\}$
	Path count	$\{\lvert P' \subseteq P(u,v) \rvert \mid \forall p_{u,v} \in P', len(p_{u,v}) = c\}$

Answer Type	Task	Description
Bool	Linked by edge	$\{\phi_E(u,v) \in T'_E\}$
	Has path	$\{P(u,v) \neq \emptyset\}$
Path	Find paths	$\{P' \subseteq P(u,v) \mid len(p_{u,v}) = c, p_{u,v} \in P'\}$
	Shortest path	$\{p'_{u,v} \in P(u,v) \mid len(p'_{u,v}) = \min(len(p_{u,v}) \mid p_{u,v} \in P(u,v))\}$
Graph	Ego graph	$\{(v_1,v_2) \in E \mid d(u,v_1) <= c, d(u,v_2) <= c\}$
Link Prediction	Link Prediction	$(E', T'_E) \rightarrow \{0,1\}$

- Unseen subtasks: trained to find author neighbors → tested to find venue neighbors
  trained to count 1-hop neighbors → tested to count 2-hop neighbors

Investigating Instruction Tuning Large Language Models on Graphs. COLM 2024.

Examined Answer Types, Tasks, and Subtasks

Answer Type	Task	Description
Node	Find neighbors	$\{v \in N(u) \mid \phi_E(u,v) \in T'_E\}$
	Nodes shared neighbors	$\{v \mid \forall t_e \in T'_E,\ \exists w \in V,\ \phi_E(u,w) = \phi_E(v,w) = t_e\}$
	N-hop neighbors	$\{v \mid \phi_V(v) \in T'_V,\ d(u,v) <= c\}$
Pair	Find pairs	$\{(v_1,v_2) \mid \phi(v_1,v_2) \in T'_E, u \in \{v_1,v_2\}\}$
	Pairs shared neighbors	$\{(v_1,v_2) \mid \exists W \subset V : \forall w \in W, \phi_E(v_1,w) = \phi_E(v_2,w) \in T'_E \wedge \lvert W\rvert = c\}$
Count	Degree count	$\{\lvert V'\rvert \mid V' \subseteq N(u) : \forall v \in V', \phi_E(u,v) \in T'_E\}$
	Node count within N-hop	$\{\lvert V'\rvert \mid V' \subset V : \forall v \in V', d(u,v) <= c\}$
	Path count	$\{\lvert P'\rvert \subseteq P(u,v) \mid \lvert\forall p_{u,v} \in P', len(p_{u,v}) = c\}$

Answer Type	Task	Description
Bool	Linked by edge	$\{\phi_E(u,v) \in T'_E\}$
	Has path	$\{P(u,v) \neq \emptyset\}$
Path	Find paths	$\{P' \subseteq P(u,v) \mid len(p_{u,v}) = c, p_{u,v} \in P'\}$
	Shortest path	$\{p'_{u,v} \in P(u,v) \mid len(p'_{u,v}) = \min(len(p_{u,v}) \mid p_{u,v} \in P(u,v))\}$
Graph	Ego graph	$\{(v_1,v_2) \in E \mid d(u,v_1) <= c, d(u,v_2) <= c\}$
Link Prediction	Link Prediction	$(E', T'_E) \rightarrow \{0,1\}$

- Unseen answer types: trained to find author neighbors (answer type: node) → tested to predict if two papers are connected via paths (answer type: bool)

Investigating Instruction Tuning Large Language Models on Graphs. COLM 2024.

Datasets and Models

- Datasets
  - MAPLE [1] – Node: paper, author, venue
  - Amazon [2] – Node: product, brand, category
- Models (instruction/chat versions)
  - Llama-2 7B
  - Mistral 7B
  - Gemma 7B

https://github.com/yuzhimanhua/MAPLE

README License

Dataset Statistics

Folder	Field	#Papers	#Labels	#Venues	#Authors	#References
Art	Art	58,373	1,990	98	54,802	115,343
Philosophy	Philosophy	59,296	3,758	98	36,619	198,010
Geography	Geography	73,883	3,285	98	157,423	884,632
Business	Business	84,858	2,392	97	100,525	685,034
Sociology	Sociology	90,208	1,935	98	85,793	842,561
History	History	113,147	2,689	99	84,529	284,739
Political_Science	Political Science	115,291	4,990	98	93,393	480,136
Environmental_Science	Environmental Science	123,945	694	100	265,728	1,217,268
Economics	Economics	178,670	5,205	97	135,247	1,042,253

[1] *The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study*. WWW 2023.
[2] *Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects*. EMNLP 2019.

Performance of Graph Instruction Tuning

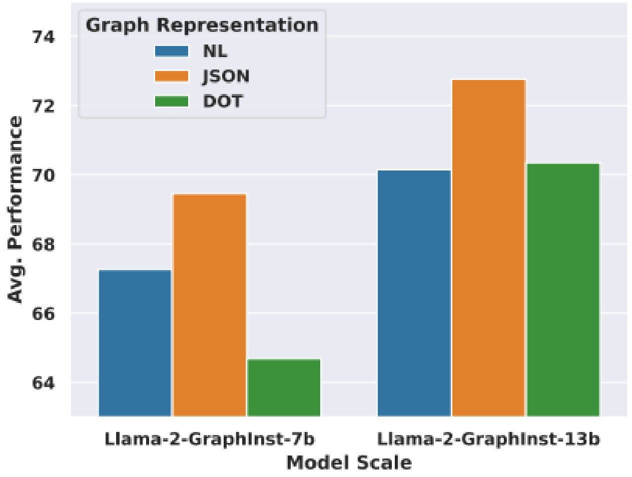
	Amazon								Maple							
	Node	Pair	Count	Bool	Path	Graph	LP	AVG	Node	Pair	Count	Bool	Path	Graph	LP	AVG
	Baselines															
Llama-2-chat$_{NL}$	1.97	**2.08**	0.00	62.83	**15.39**	10.98	41.79	**12.91**	2.85	**3.26**	**0.16**	58.91	**16.21**	18.96	47.33	**14.56**
Llama-2-chat$_{JSON}$	2.16	2.04	0.00	62.83	3.87	7.78	41.79	11.48	**4.12**	1.90	0.00	58.91	8.32	10.68	47.33	12.98
Llama-2-chat$_{DOT}$	**2.48**	1.66	0.00	62.83	1.56	**12.78**	41.79	11.78	2.31	2.34	0.00	58.91	4.42	**22.16**	47.33	13.36
Mistral-Inst$_{NL}$	0.01	3.98	**12.89**	**37.55**	18.15	7.27	58.21	13.84	0.03	5.81	14.04	**42.98**	20.54	6.70	52.67	15.04
Mistral-Inst$_{JSON}$	**2.91**	**8.38**	12.43	37.30	12.29	8.86	58.21	**14.75**	**4.24**	**10.81**	**14.45**	41.09	9.74	7.74	52.67	**15.77**
Mistral-Inst$_{DOT}$	1.65	5.28	8.00	37.17	**18.43**	**12.50**	58.21	14.02	3.07	4.85	12.41	41.09	**23.34**	**9.72**	52.67	15.74
Gemma-Inst$_{NL}$	13.85	25.19	2.59	**65.92**	**34.64**	28.85	44.96	24.42	15.17	29.95	3.53	**72.75**	**34.65**	23.03	**48.72**	26.35
Gemma-Inst$_{JSON}$	15.50	26.54	**6.67**	65.14	30.00	29.45	36.00	24.74	8.51	26.45	**8.42**	65.61	29.32	22.19	45.59	23.50
Gemma-Inst$_{DOT}$	**16.83**	**35.91**	4.06	64.76	31.19	**37.69**	**60.33**	**28.72**	**18.17**	**34.76**	2.38	64.99	29.22	**36.98**	42.86	**27.25**
	Finetuned															
Llama-2-GraphInst$_{NL}$	74.34	65.97	45.76	93.57	**55.16**	64.83	85.47	67.26	73.28	67.78	44.05	96.13	53.65	77.69	67.07	66.74
Llama-2-GraphInst$_{JSON}$	**80.20**	**68.49**	**46.48**	**96.48**	52.75	**65.39**	**85.02**	**69.46**	**75.33**	**68.42**	46.62	**98.11**	**55.21**	**80.06**	64.42	**68.30**
Llama-2-GraphInst$_{DOT}$	73.64	64.26	43.76	91.20	50.07	61.39	77.11	64.69	70.59	63.69	**47.32**	94.35	54.47	74.83	**69.08**	65.86
Mistral-GraphInst$_{NL}$	87.43	75.38	48.47	98.13	66.55	80.09	75.23	75.16	**86.17**	76.39	48.86	97.80	68.55	**86.14**	76.97	**75.69**
Mistral-GraphInst$_{JSON}$	**89.63**	**81.18**	**50.77**	**98.73**	62.16	**83.32**	76.15	**77.11**	82.96	**79.58**	**50.94**	**99.16**	**68.61**	84.13	75.95	75.64
Mistral-GraphInst$_{DOT}$	86.30	72.91	46.24	96.97	**68.71**	81.01	**77.98**	74.44	79.01	74.96	50.74	98.95	66.47	85.69	**79.02**	74.03
Gemma-GraphInst$_{NL}$	87.34	76.30	46.51	97.36	**68.06**	77.31	**85.17**	75.43	**88.90**	74.76	47.84	97.49	61.50	86.04	**77.53**	75.20
Gemma-GraphInst$_{JSON}$	**90.15**	**78.11**	**49.98**	**99.23**	65.68	78.08	82.42	**76.98**	88.50	**75.33**	**51.74**	**98.64**	63.39	83.15	70.91	**75.50**
Gemma-GraphInst$_{DOT}$	87.43	78.09	47.96	96.36	67.43	**83.50**	83.94	76.37	85.75	74.00	50.14	98.43	**68.14**	**88.71**	70.77	75.29

Which representation of the graph is the best?

- Scalability: Natural language has the most compact representation and can handle the largest graph in a limited context budget.

	Amazon			Maple		
	# Avg Tokens	# Max Nodes	# Max Edges	# Avg Tokens	# Max Nodes	# Max Edges
NL	1869.56	226	324	1033.61	280	326
JSON	1972.44	199	289	1161.03	277	321
DOT	2011.01	192	288	1181.22	277	321

- Performance: The JSON format yields the best overall performance for all three models.



Subtask Generalization

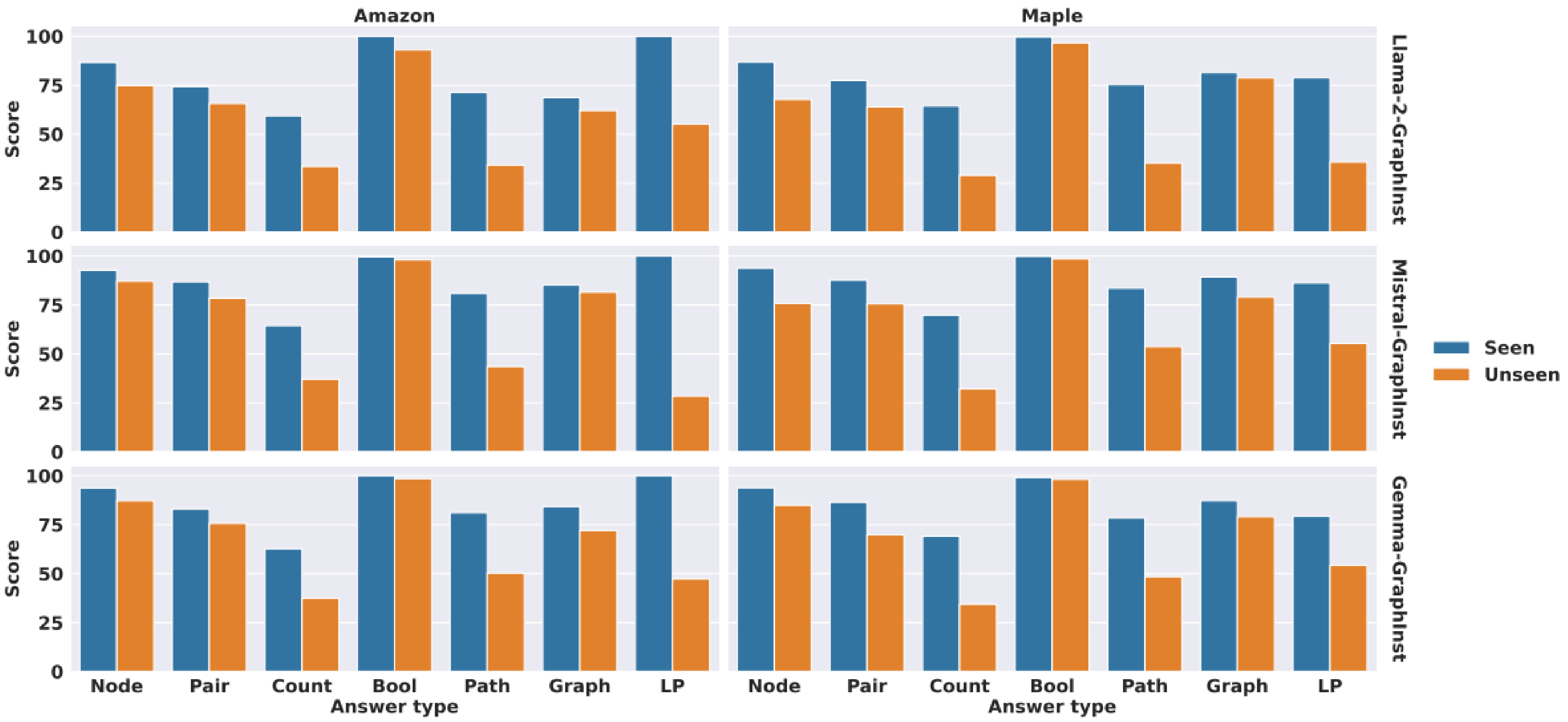


Figure 3: Experiment results of sub-task generalization on two datasets.

Answer Type Generalization & Domain Generalization

	Amazon					
	Node	**Pair***	**Count**	**Bool***	**Path**	**Graph***
Mistral-Inst$_{JSON}$	2.91	8.38	12.43	37.30	12.29	8.86
Mistral-GraphInst$_{JSON}$	89.63	81.18	50.77	98.73	62.16	83.32
Mistral-GraphInst-masked$_{JSON}$	88.09	56.43	49.91	90.18	59.31	53.65
	Maple					
Mistral-Inst$_{JSON}$	4.24	10.81	14.45	41.09	9.74	7.74
Mistral-GraphInst$_{JSON}$	82.96	79.58	50.94	99.16	68.61	84.13
Mistral-GraphInst-masked$_{JSON}$	79.70	40.90	50.48	77.15	64.30	36.64

Table 5: Answer Type Generalization, where tasks *Pair*, *Bool* and *Graph* are unseen when training *Mistral-GraphInst-masked*.

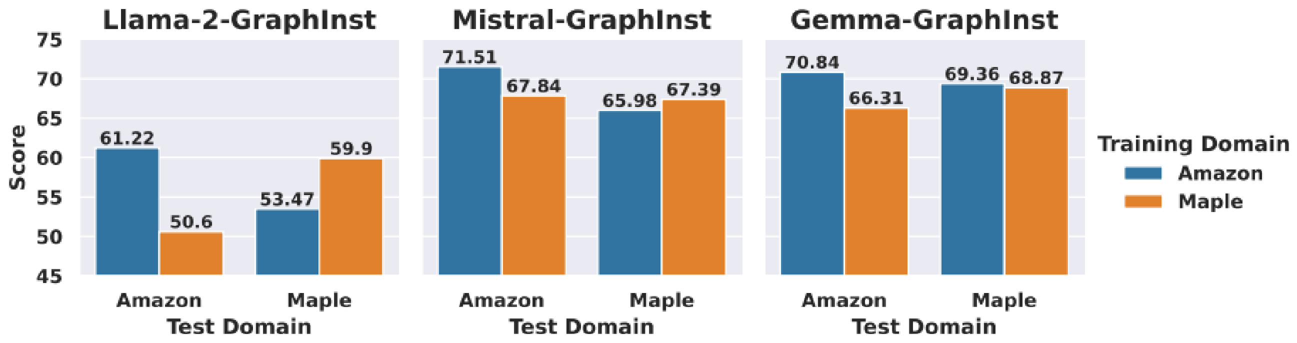


Figure 4: Compare LLMs of different scales on domain generalization.

Take-Away Messages

- Different graph representations: Natural language has the most compact representation; JSON yields the best performance after tuning.
- Different levels of generalization: LLMs could be overfitted by signals seen during training and hard to generalize to unseen subtasks, answer types, and domains.

- Limitations
  - No studies on task-level (i.e., the level between subtasks and answer types) generalization.
    - E.g., trained on degree counting → tested on path counting
  - No studies on subdomain-level generalization
    - E.g., trained on the academic graph of CS papers → tested on the graph of medicine papers

Midterm Project Presentations

- 5 groups
- Each group has 10 minutes for presentation and 3 minutes for Q&A.
  - The number of presenters per group is not limited.

- If you would like to use the instructor’s laptop, please send me the slides via email at least 30 minutes before the lecture.

- Presentation order: Last name in alphabetical order
  - 1. Hasnat and Rithik
  - 2. Shaohuai
  - 3. Omnia and Michael
  - 4. Yichen and Ethan
  - 5. Shuo and Hangxiao

Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html