Part III: Text Mining with Structured Information

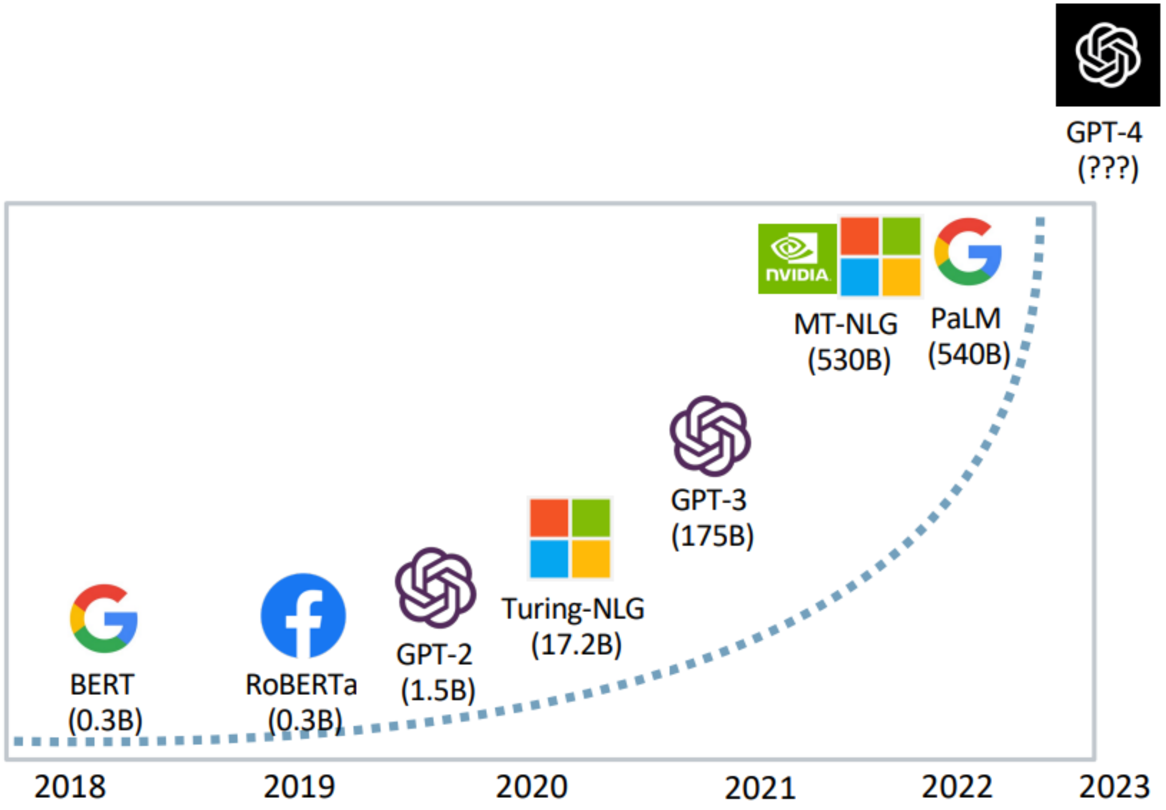
Bowen Jin, Yu Zhang, Sha Li, Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign

Mar 4, 2024

Tutorial Website:

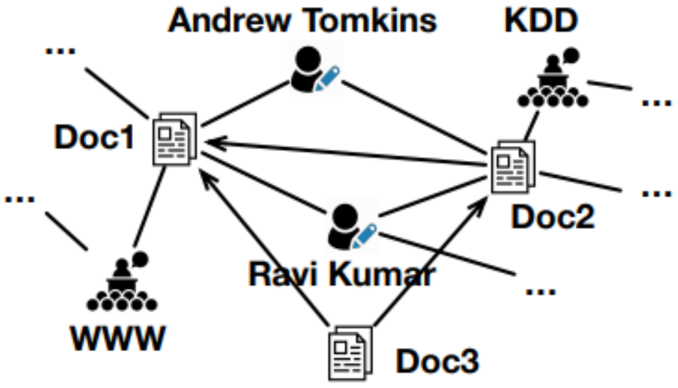
Pre-trained Language Models (PLMs) for Text Mining

- A unified model to perform different text mining tasks with a few or zero examples
  - I went to the zoo to see giraffes, lions, and {zebras, spoon}. *(Lexical semantics)*
  - I was engaged and on the edge of my seat the whole time. The movie was {good, bad}. *(Sentiment analysis)*
  - The word for "pretty" in Spanish is {bonita, hola}. *(Translation)*
  - 3 + 8 + 4 = {15, 11} *(Math)*
  - …

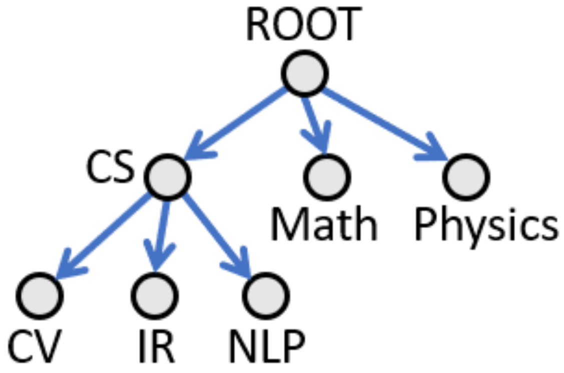


- Are PLMs aware of structured information?

Structured Information Associated with Text



Metadata/Network



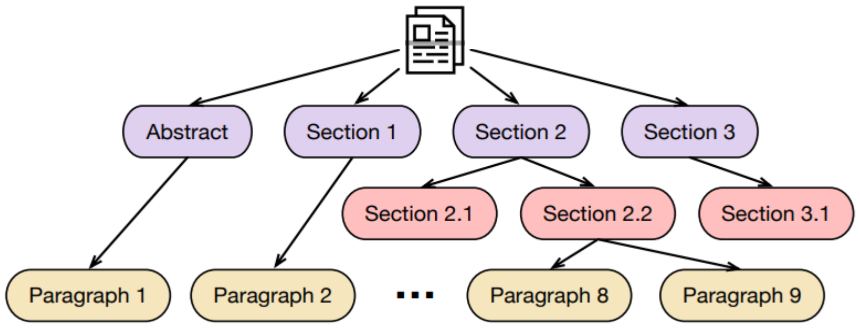
Hierarchy/Taxonomy

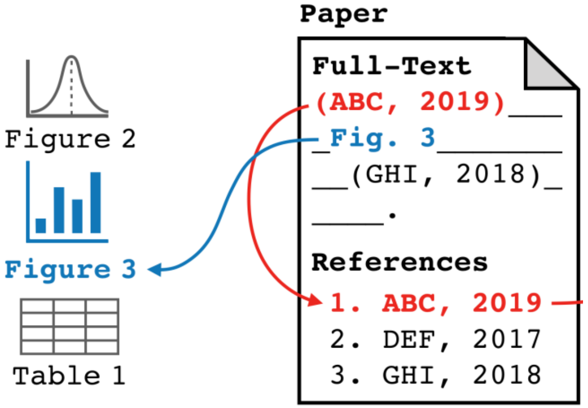
WIKIDATA

ctd

DBpedia

Knowledge Graph





In-Document Structure

PLMs may not be structure-aware!

You

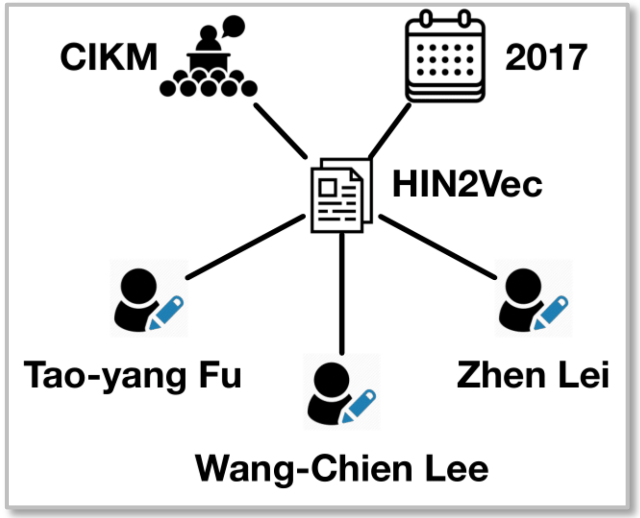
What is the most cited paper in WWW 2017? Please list its title and authors.

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.

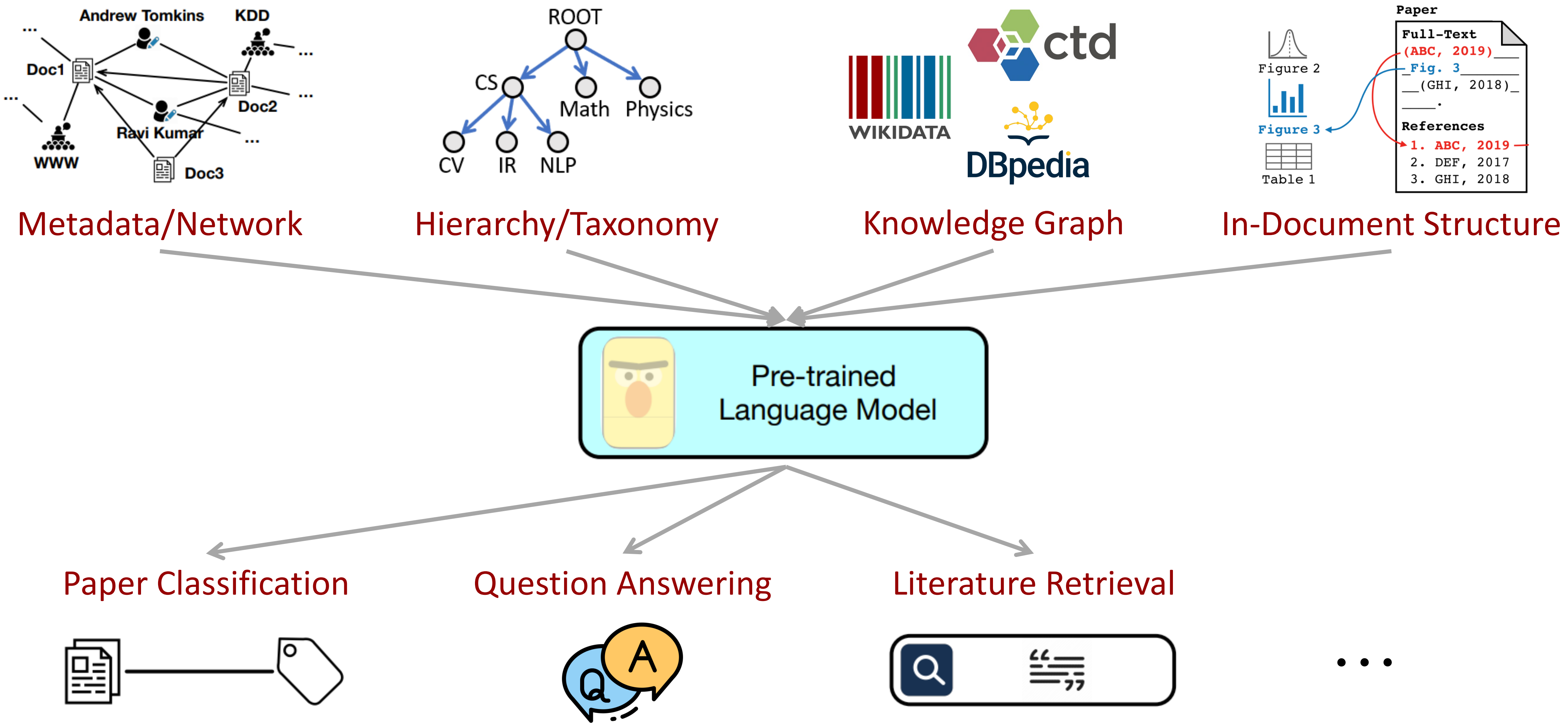


HIN2Vec was published in CIKM 2017.

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

ChatGPT 3.5, queried on Jan 23, 2024

Text Mining with Structured Information



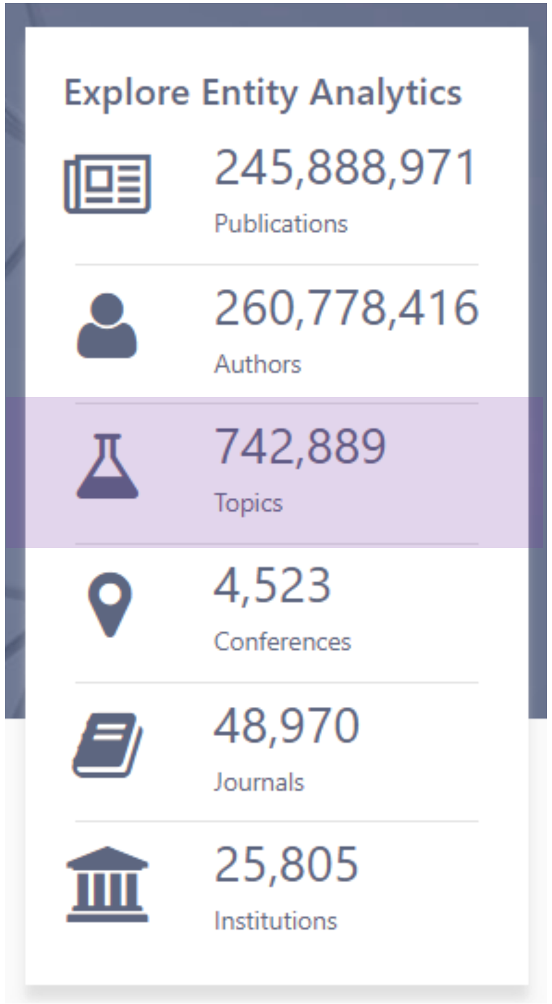
Outline

- Structure-enhanced Text Classification
  - Metadata
  - Hierarchy
- Structure-enhanced Question Answering
  - Knowledge Graph
- Structure-enhanced Language Model Pre-training
  - Citation Link
  - Integrating Multiple Types of Structured Information

Outline

- Structure-enhanced Text Classification
  - Metadata
    - Metadata as Additional Features
    - Metadata as Proximity Indicators
  - Hierarchy
- Structure-enhanced Question Answering
- Structure-enhanced Language Model Pre-training

Extremely Fine-Grained Scientific Paper Classification



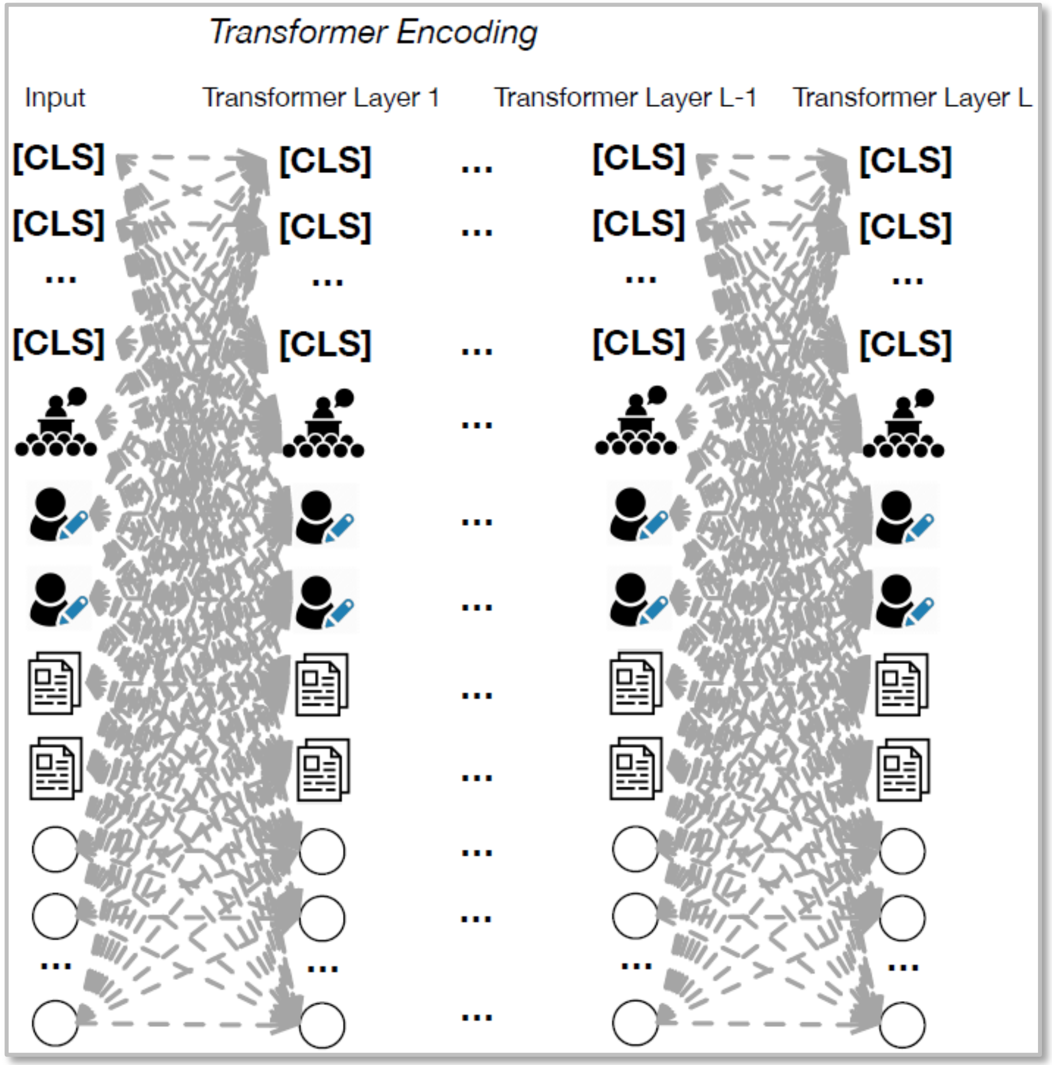
- The Microsoft Academic Graph has 740K+ categories.
- The Medical Subject Headings (MeSH) for indexing PubMed papers contain 30K+ categories.
- Each paper can be relevant to more than one category (5-15 categories for most papers).

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, …

Metadata as Additional Features: MATCH

- How to add metadata?
  - Concatenating the [CLS] tokens, metadata instances, and text as the input into Transformer.
  - E.g., **[CLS$_1$] [CLS$_2$] … [CLS$_C$] [Venue_WWW] [Author_Andrei Broder] [Author_Ravi Kumar] … [Reference_2066636486] [Reference_1976969221] … [Word_graph] [Word_structure] [Word_in] [Word_the] [Word_web] …**
  - The fully connected attention mechanism will enable interaction between text and metadata.



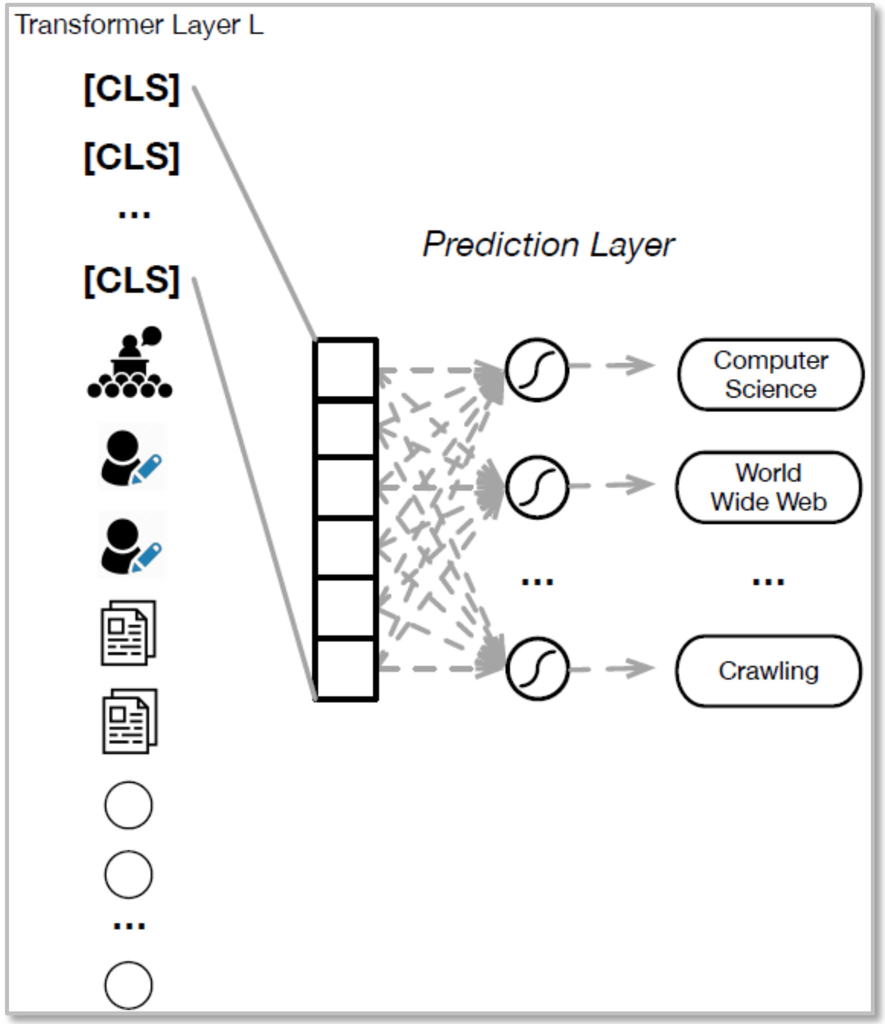
Zhang, Y., Shen, Z., Dong, Y., Wang, K. & Han, J. "MATCH: Metadata-Aware Text Classification in A Large Hierarchy", WWW'21.

Metadata as Additional Features: MATCH

- The final layer is then connected to the sigmoid functions corresponding to all labels. The output of the l-th sigmoid function (π_{dl}) denotes the probability that document d should be tagged with label l.

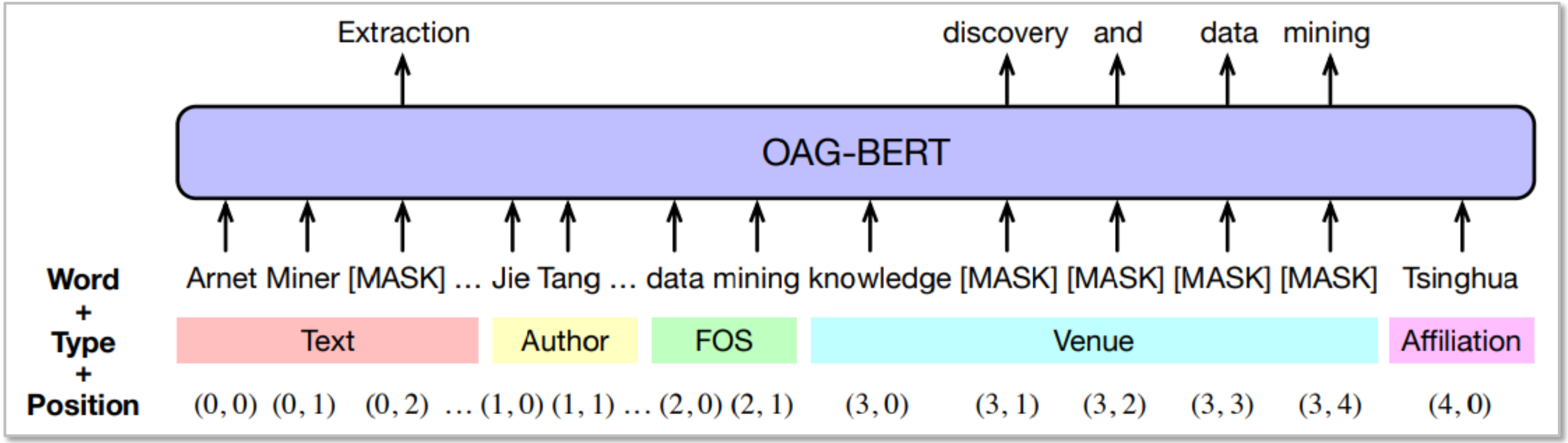
- The model is trained by minimizing the cross-entropy loss:

$$-\sum_{d \in \mathcal{D}} \sum_{l \in \mathcal{L}} \left(y_{dl} \log \pi_{dl} + (1 - y_{dl}) \log(1 - \pi_{dl}) \right),$$



Metadata as Additional Features: OAG-BERT

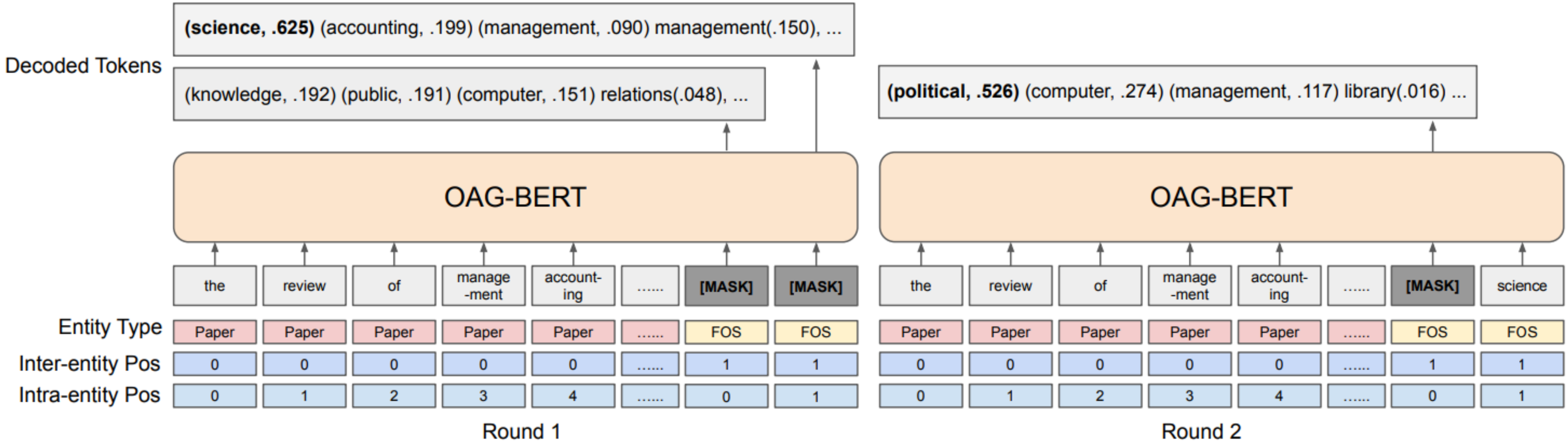
- Heterogeneous entity type embedding makes the model aware of different metadata types.
- Span-aware entity masking selects a continuous span within long entities (e.g., the venue “knowledge discovery and data mining”).
- 2-dimensional positional embedding jointly models inter and intra-entity token orders.



Liu, X., Yin, D., Zheng, J., Zhang, X., Zhang, P., Yang, H., Dong, Y., & Tang, J. "OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services", KDD’22.

Metadata as Additional Features: OAG-BERT

- Classification via probing



Outline

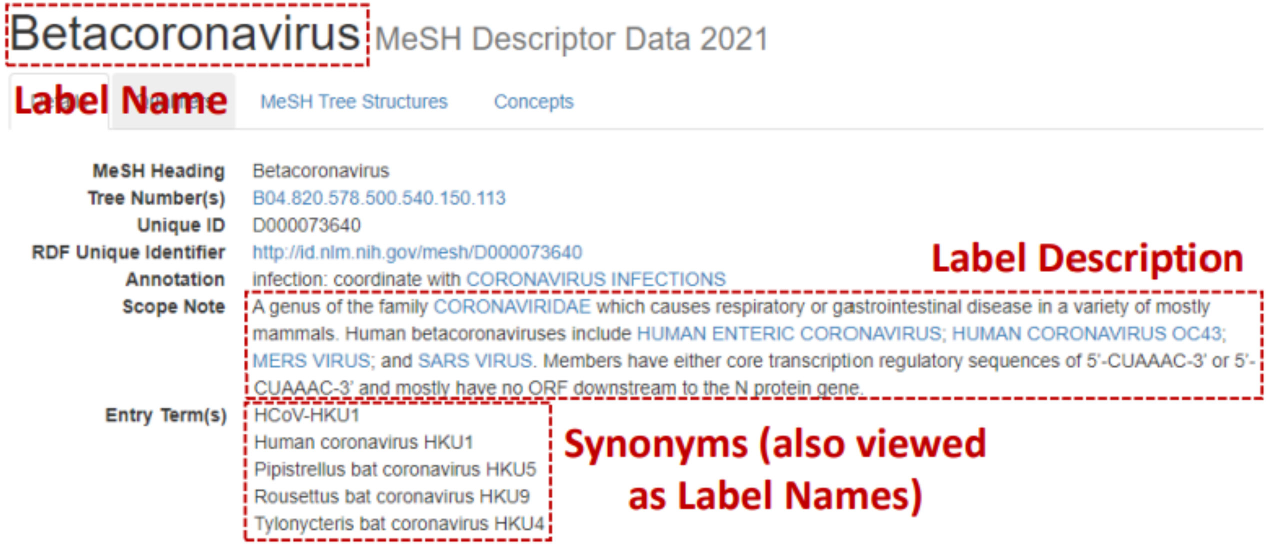
- Structure-enhanced Text Classification
  - Metadata
    - Metadata as Additional Features
    - Metadata as Proximity Indicators
  - Hierarchy
- Structure-enhanced Question Answering
- Structure-enhanced Language Model Pre-training

Classification as Predicting Proximity between Paper and Label

- Labels also have text information.
  - Label name
  - Synonyms (optional)
  - Definition/description (optional)

- A naïve classification approach:
  - Use a PLM to encode each paper
  - Use the same PLM to encode each label (described by all available text information)
  - Find the nearest label neighbors for each paper
  - Not performing well if the PLM is unfine-tuned!

(a) Label "Webgraph" from Microsoft Academic (https://academic.microsoft.com/topic/2777569578/).



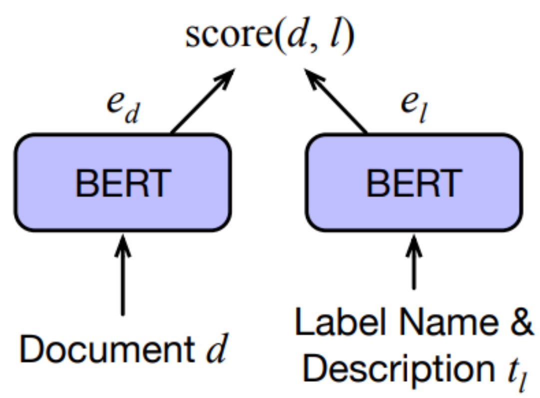
(b) Label "Betacoronavirus" from PubMed (https://meshb.nlm.nih.gov/record/ui?ui=D000073640).

If we could have some training data …

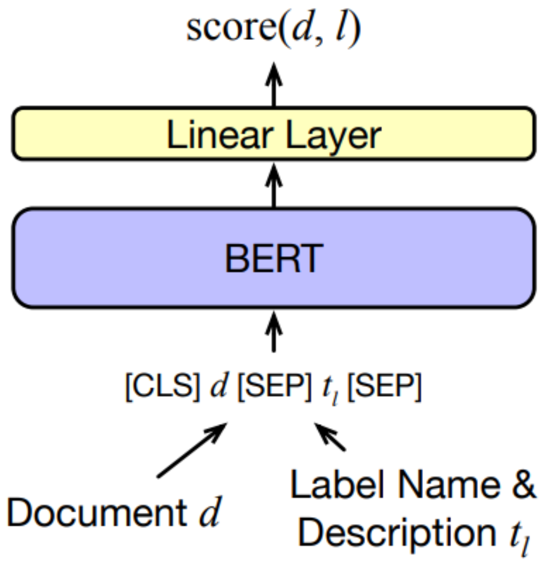
- We could use relevant (paper, category) pairs to fine-tune a pre-trained language model.
- Both Bi-Encoder and Cross-Encoder are applicable.

Should be large

Bi-Encoder



Should be large



Cross-Encoder

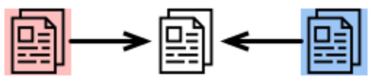
- However, human-annotated training samples are NOT available in many cases!
  - We are asking annotators to find ~10 relevant categories from ~100,000 candidates!

Using Metadata Information 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.
  - …

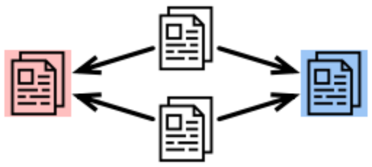
- The notion of meta-paths and meta-graphs

(a) meta-path: PAP

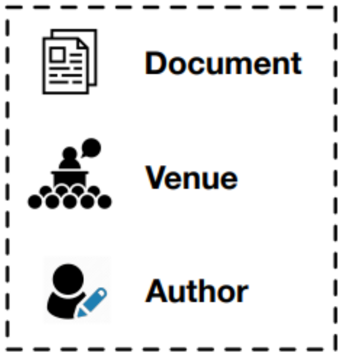


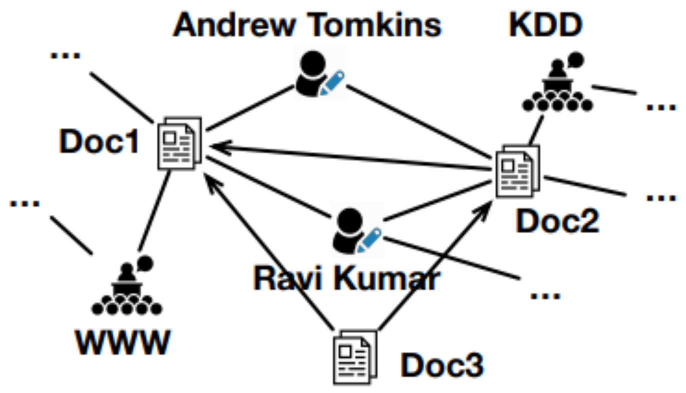
(b) meta-path: P->P<-P

(c) meta-graph: P(AV)P



(d) meta-graph: P<-(PP)->P



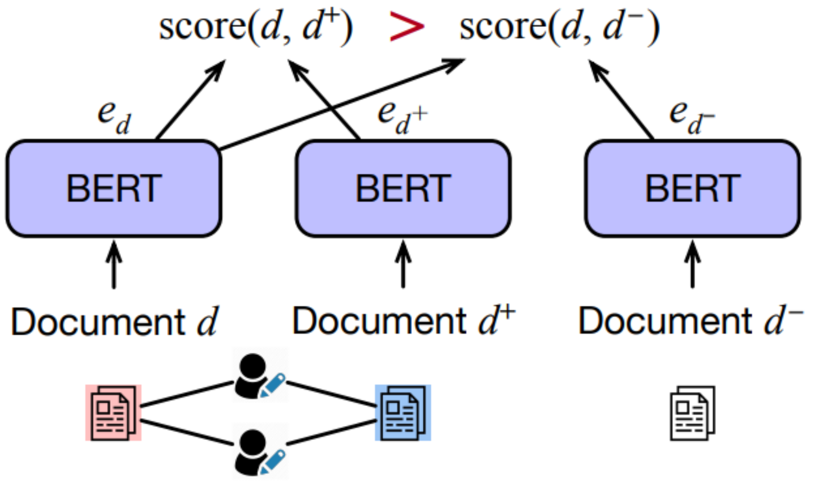


Metadata-Induced Contrastive Learning

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

Bi-Encoder

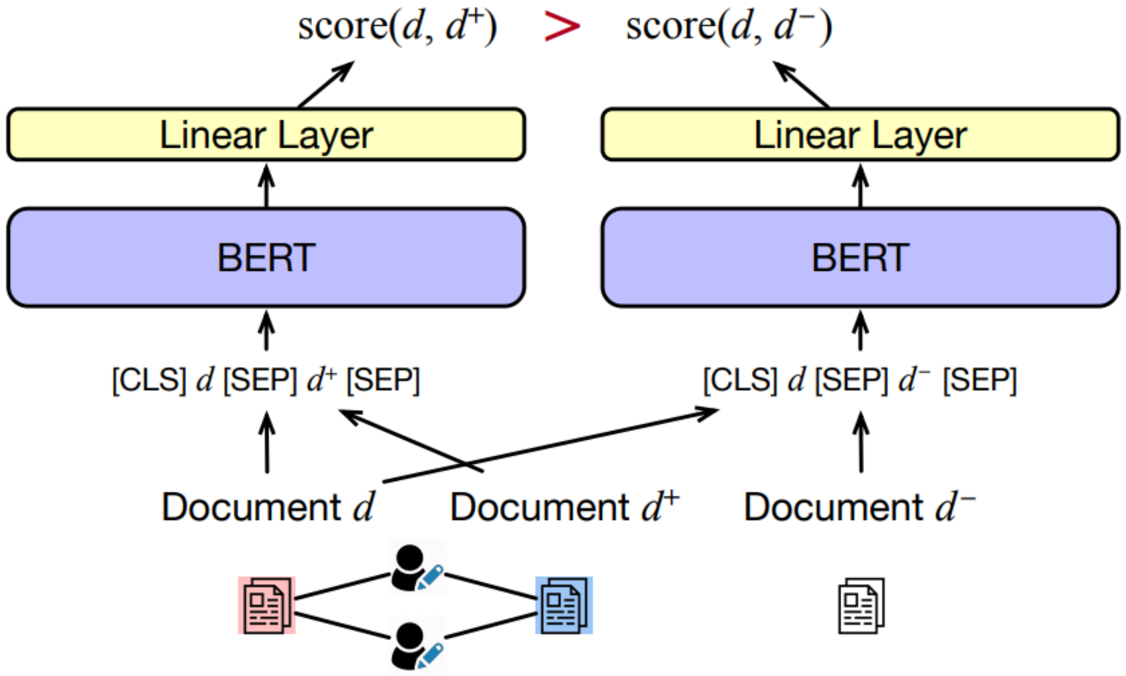
Should be larger Should be smaller



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

Cross-Encoder

Should be larger Should be smaller



Zhang, Y., Shen, Z., Wu, C., Xie, B., Wang, Y., Wang, K. & Han, J. "Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification", WWW'22.

MICoL: Experimental Results

- MICoL significantly outperforms scientific PLMs, zero-shot text classification baselines, and text-based contrastive learning baselines.
- MICoL is competitive with the supervised SOTA trained on 10K–50K labeled documents.

	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
Supervised	MATCH [68] (10K Training)	0.4423**	0.2851**	0.2152**	0.3375**	0.3003**	0.6915	0.3869*	0.2785**	0.4649	0.3896
	MATCH [68] (50K Training)	0.6215**	0.4280**	0.3269**	0.4987**	0.4489**	0.7701	0.4716	0.3585	0.5497	0.4750
	MATCH [68] (100K Training)	0.8321	0.6520	0.5142	0.7342	0.6761	0.8286	0.5680	0.4410	0.6405	0.5626
	MATCH [68] (Full, 560K+ Training)	0.9114	0.7634	0.6312	0.8486	0.8076	0.9151	0.7425	0.6104	0.8001	0.7310

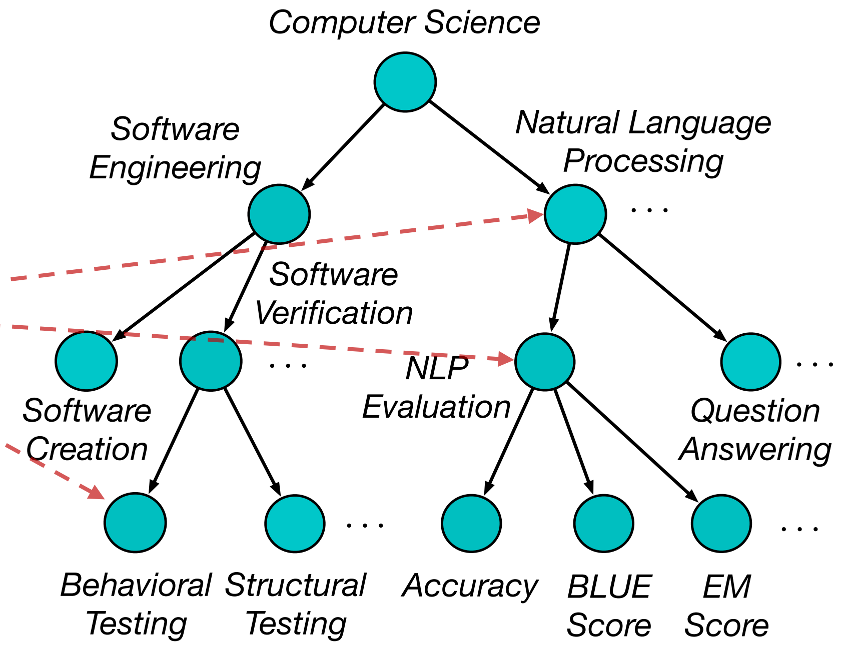
Outline

- Structure-enhanced Text Classification
  - Metadata
  - Hierarchy
    - Hierarchy for Label Space Pruning
    - Hierarchy for Label Relationship Learning
- Structure-enhanced Question Answering
- Structure-enhanced Language Model Pre-training

Weakly-supervised Hierarchical Multi-Label Text Classification

- The taxonomy is a directed acyclic graph (DAG).
- Each paper can have multiple categories distributed on different paths.
- Category names can be phrases and may not appear in the corpus.

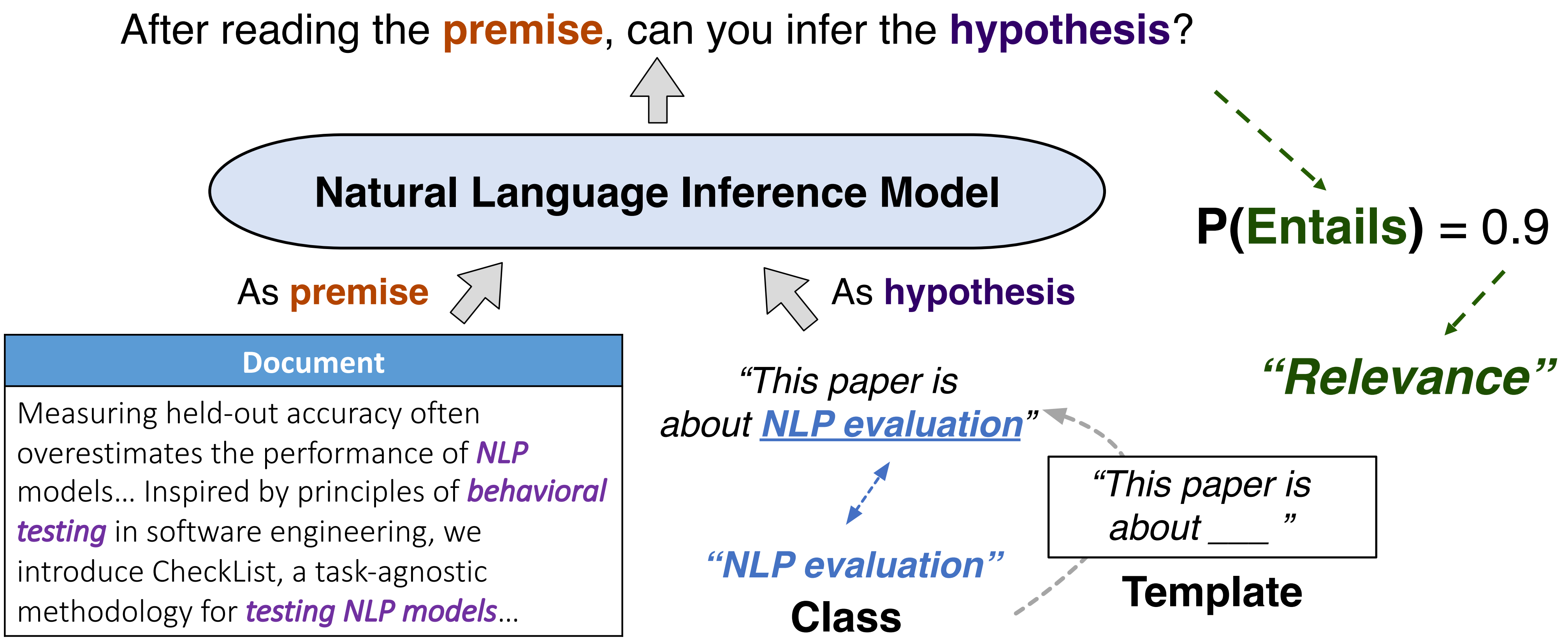
Document
Measuring held-out accuracy often overestimates the performance of ***NLP*** models… Inspired by principles of ***behavioral testing*** in software engineering, we introduce CheckList, a task-agnostic methodology for ***testing NLP models***…



Shen, J., Qiu, W., Meng, Y., Shang, J., Ren, X., & Han, J., "TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names", NAACL'21.

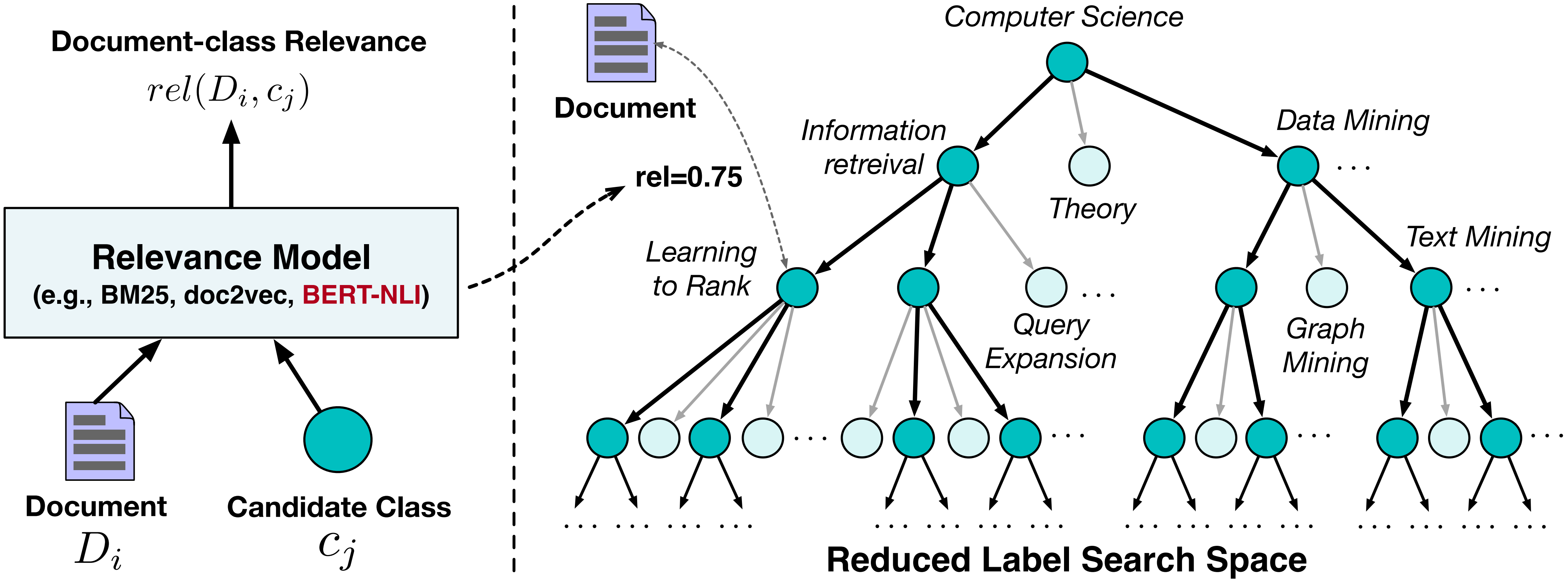
TaxoClass: Document-Class Relevance Calculation

- ❑ How to use the knowledge from pre-trained LMs?
- ❑ Relevance model: BERT/RoBERTa fine-tuned on the NLI task
  - ❑ https://huggingface.co/roberta-large-mnli



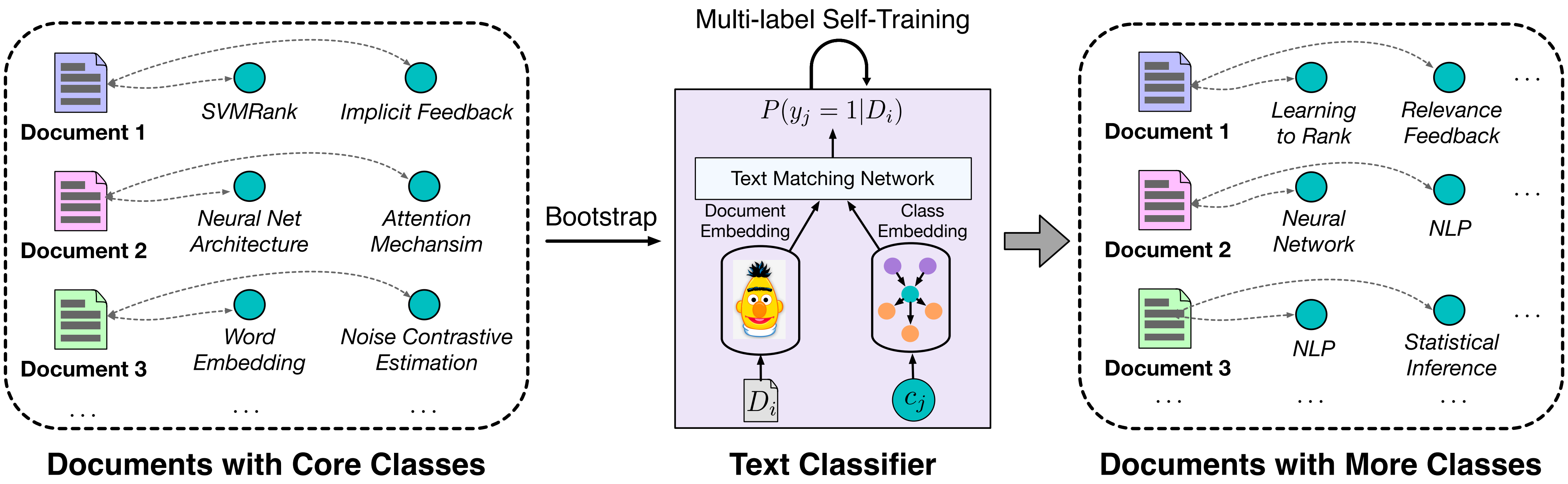
TaxoClass: Top-Down Exploration

- How to use the taxonomy?
- Shrink the label search space with top-down exploration
  - Use a relevance model to filter out completely irrelevant classes



TaxoClass: Identify Core Classes and More Classes

- Identify document core classes in reduced label search space
- Generalize from core classes with bootstrapping and self-training



TaxoClass: Experimental Results

Methods	Amazon		DBPedia	
	Example-F1	P@1	Example-F1	P@1
WeSHClass (Meng et al., AAAI'19)	0.246	0.577	0.305	0.536
SS-PCEM (Xiao et al., WebConf'19)	0.292	0.537	0.385	0.742
Semi-BERT (Devlin et al., NAACL'19)	0.339	0.592	0.428	0.761
Hier-0Shot-TC (Yin et al., EMNLP'19)	0.474	0.714	0.677	0.787
TaxoClass	**0.593**	**0.812**	**0.816**	**0.894**

Weakly-supervised multi-class classification method (WeSHClass)

Semi-supervised methods using 30% of training set (SS-PCEM, Semi-BERT)

Zero-shot method (Hier-0Shot-TC)

- **vs. WeSHClass**: better model document-class relevance
- **vs. SS-PCEM, Semi-BERT**: better leverage supervision signals from taxonomy
- **vs. Hier-0Shot-TC**: better capture domain-specific information from core classes

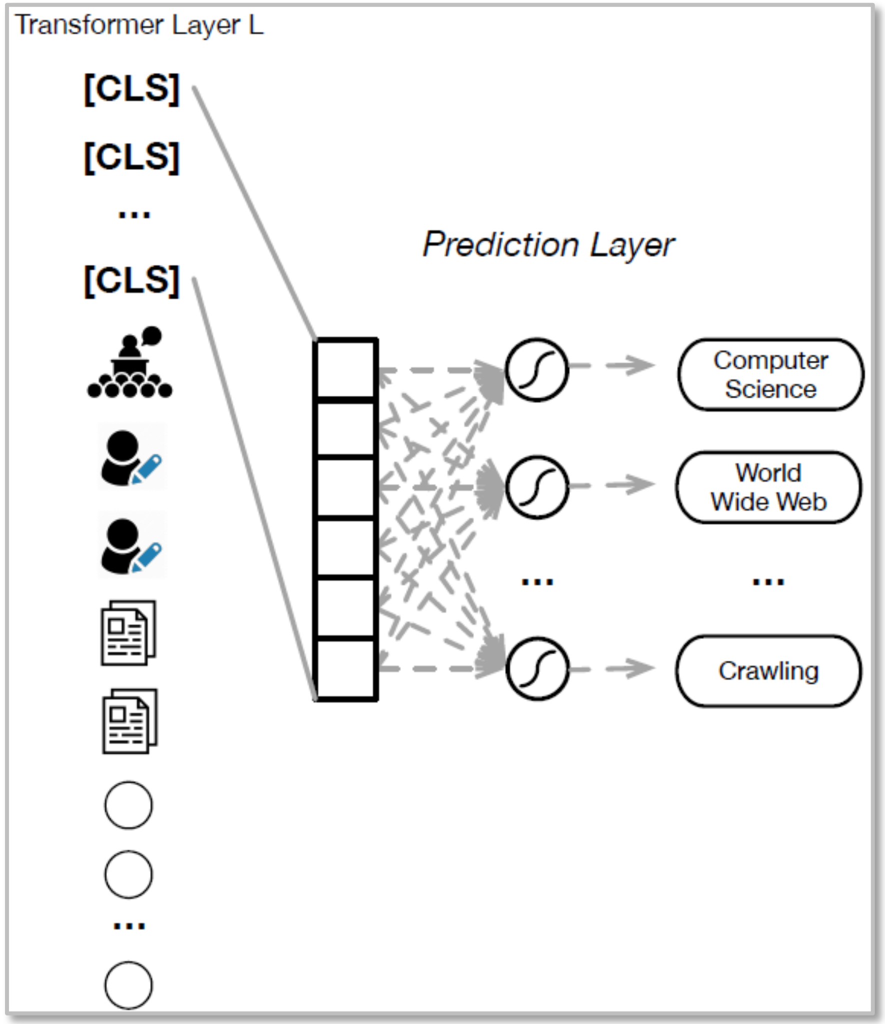
Example-F1 = $\frac{1}{N}\sum_{i=1}^{N}\frac{2|true_i \cap pred_i|}{|true_i|+|pred_i|}$, **P@1** = $\frac{\#docs\ with\ top{-}1\ pred\ dorrect}{\#total\ docs}$

Outline

- Structure-enhanced Text Classification
  - Metadata
  - Hierarchy
    - Hierarchy for Label Space Pruning
    - Hierarchy for Label Relationship Learning
- Structure-enhanced Question Answering
- Structure-enhanced Language Model Pre-training

Taxonomy-Based Regularization: Parameter Space

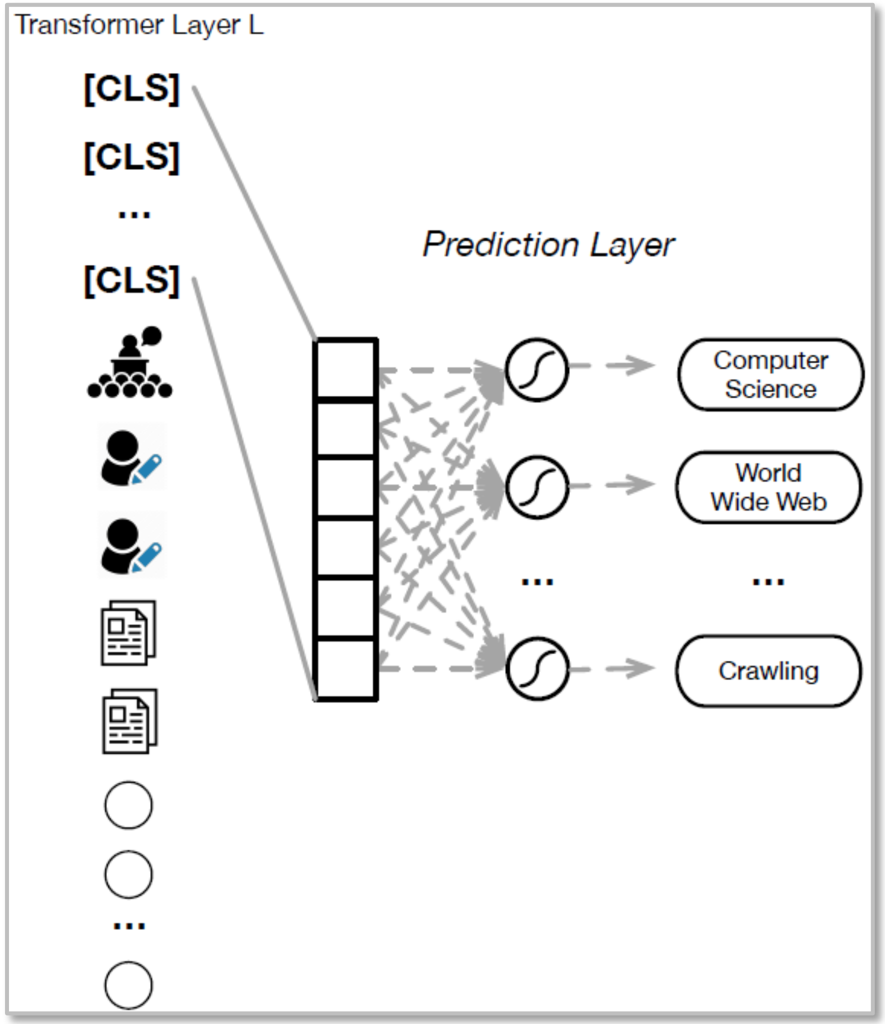
- The paper representation is connected to the sigmoid functions corresponding to all labels.
- In the parameter space, an L2-norm penalty can be adopted to enforce the parameters of each label to be similar with its parents.
  - Intuition: Judging whether a document can be tagged with "*crawling*" should bear similarities with judging whether it is related to the parent label "*world wide web*".



Gopal, S., & Yang, Y., "Recursive regularization for large-scale classification with hierarchical and graphical dependencies." KDD'13.

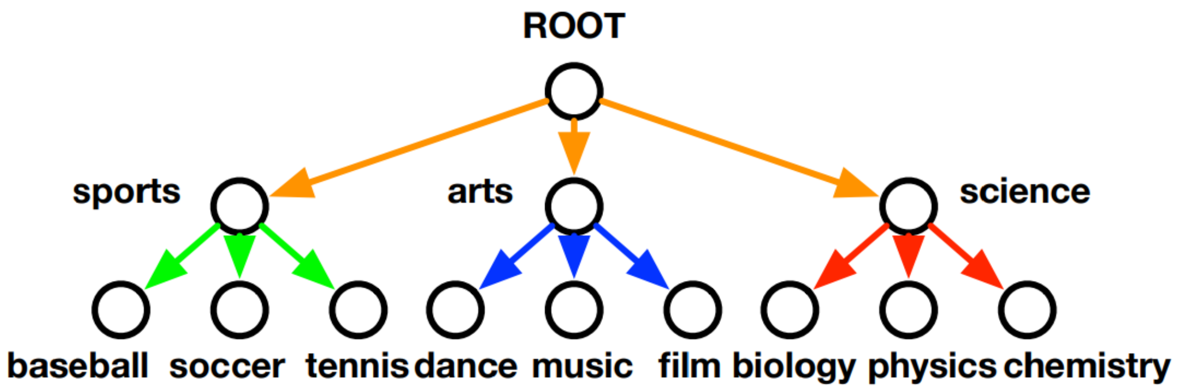
Taxonomy-Based Regularization: Output Space

- The paper representation is connected to the sigmoid functions corresponding to all labels.
- In the output space, an asymmetric relationship between parent and child labels can be modeled.
  - Intuition: If there is a 50% chance a paper will be labeled with "*crawling*", then the chance to label this paper with "*world wide web*" should be at least 50% (because the paper may be labeled with siblings of "*crawling*").



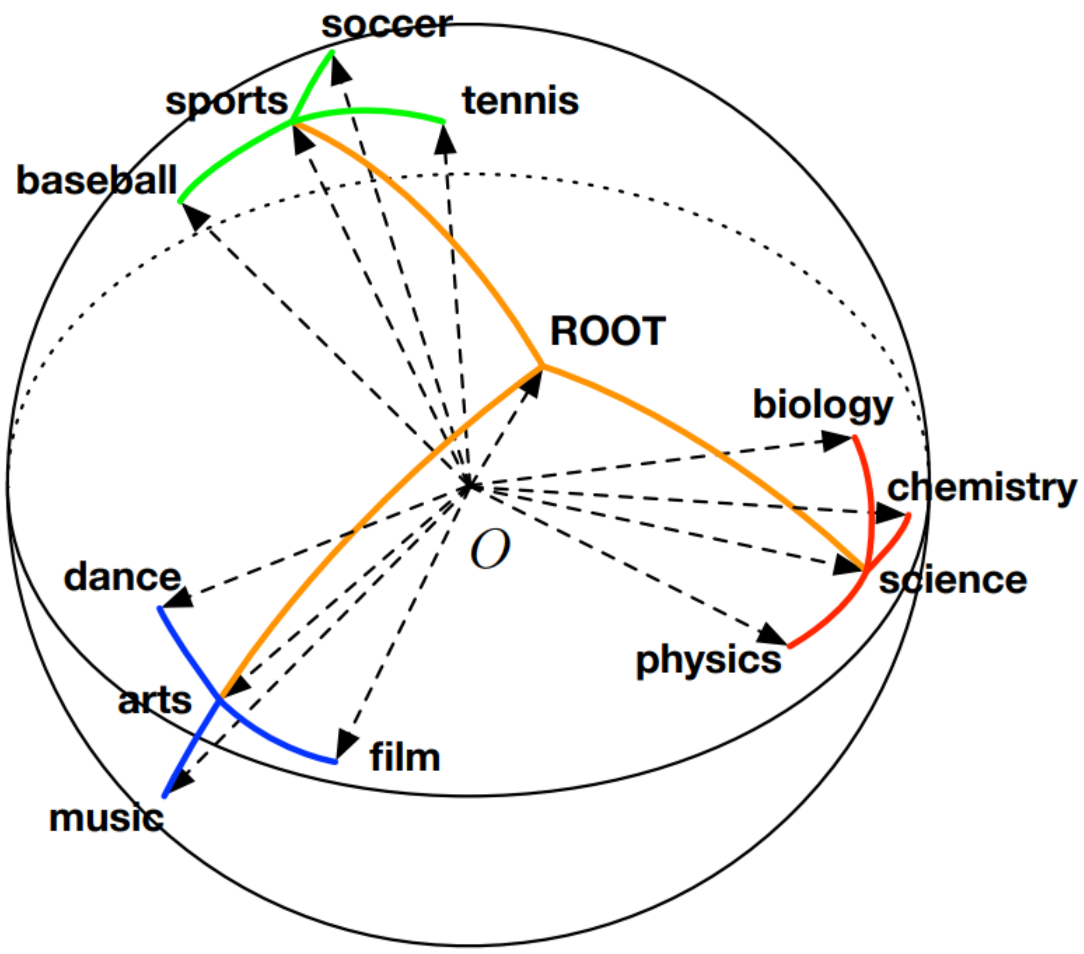
Zhang, Y., Shen, Z., Dong, Y., Wang, K. & Han, J. "MATCH: Metadata-Aware Text Classification in A Large Hierarchy", WWW'21.

Taxonomy-Based Contrastive Learning



- A category should be closer to its parent category than to its sibling categories in the embedding space.

$$\sum_{c_i \in \mathcal{T}_r \setminus \{c_r\}} \sum_{c_j \in \mathcal{T}_r \setminus \{c_r, c_i\}} \min(0, \mathbf{c}_i^\top \mathbf{c}_r - \mathbf{c}_i^\top \mathbf{c}_j - m_{\text{inter}})$$



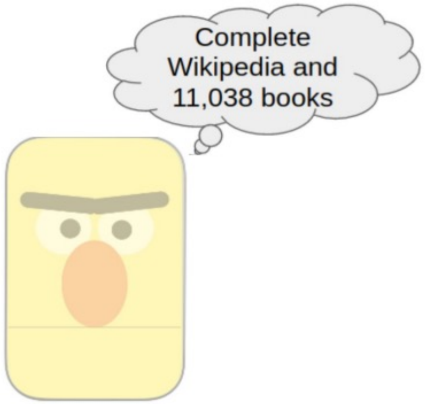
Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Zhang, C., & Han, J. "Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding", KDD'20.

Outline

- Structure-enhanced Text Classification
- Structure-enhanced Question Answering
  - Knowledge Graph
- Structure-enhanced Language Model Pre-training

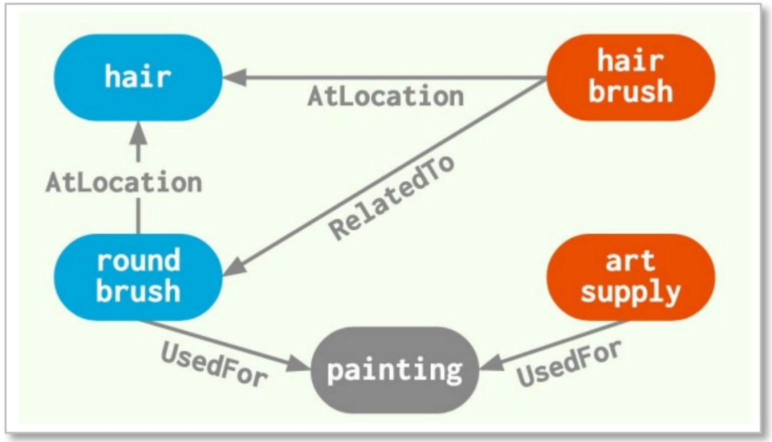
Text & KG Offer Complementary Information

- Text & Pretrained Language Model
  - Broad coverage
  - Capturing rich context



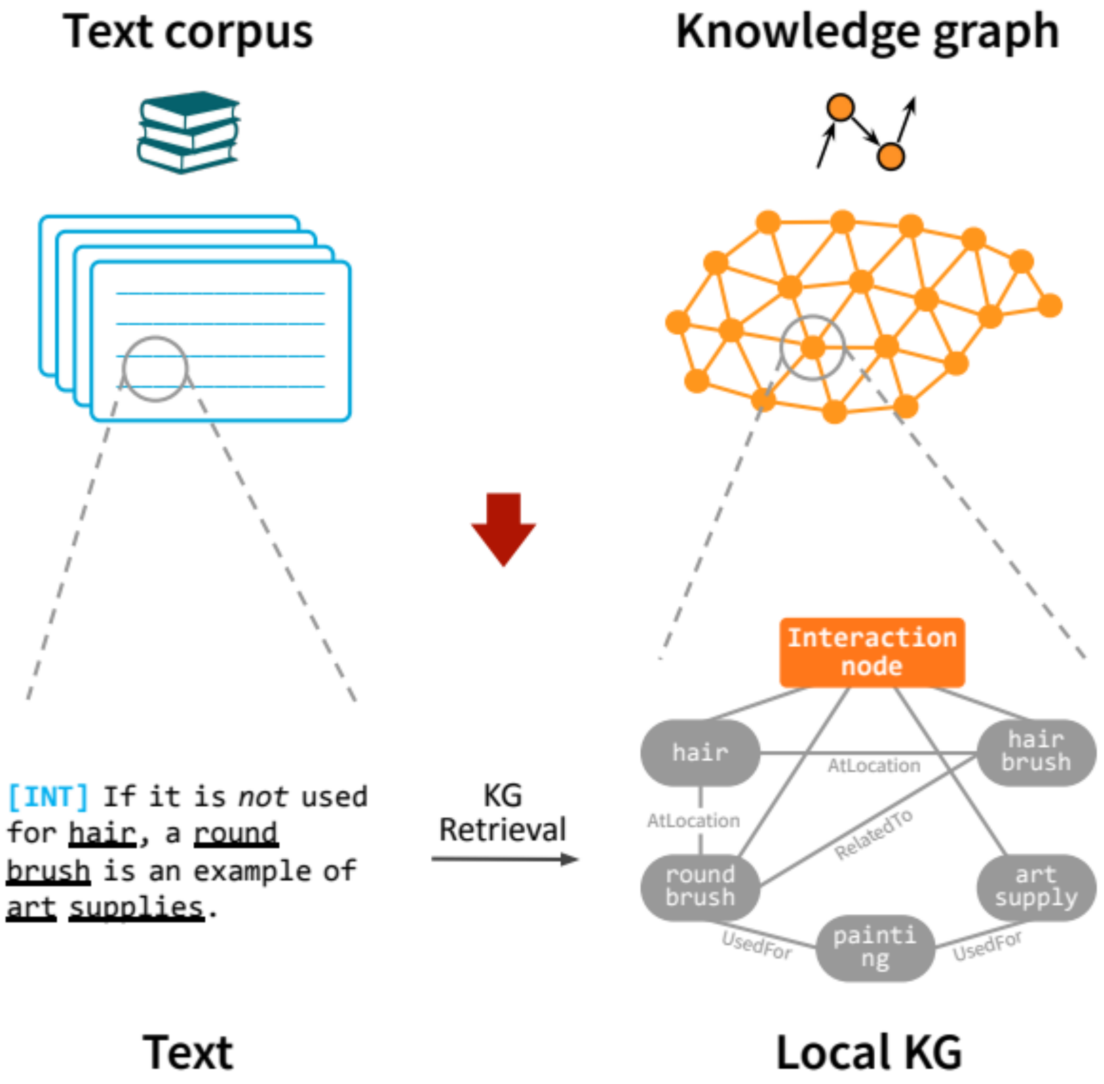
[INT] If it is *not* used for hair, a round brush is an example of art supplies.

- Knowledge Graph
  - Latent, structured relations
  - Tail knowledge
  - Multi-hop reasoning



GreaseLM: Combining Text & KG for Question Answering

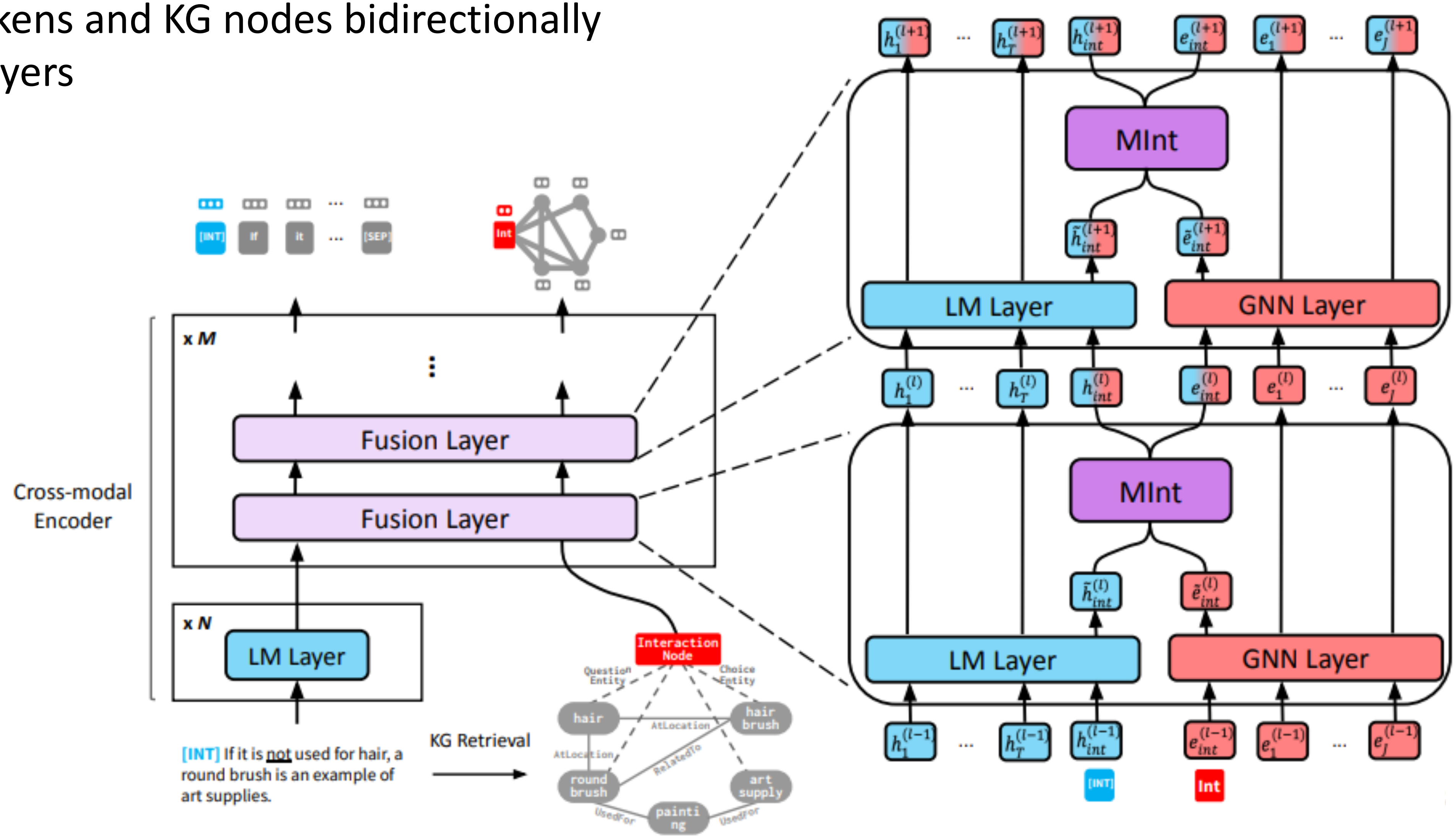
- An informative pair of (Text, Local KG) as input
  - Given a text corpus and a KG, sample a text segment and retrieve a relevant knowledge subgraph by entity linking.
  - Text can contextualize the KG.
  - KG can ground the text.



Zhang, X., Bosselut, A., Yasunaga, M., Ren, H., Liang, P., Leskovec, J., & Manning, C., "GreaseLM: Graph REASoning Enhanced Language Models for Question Answering." ICLR'22.

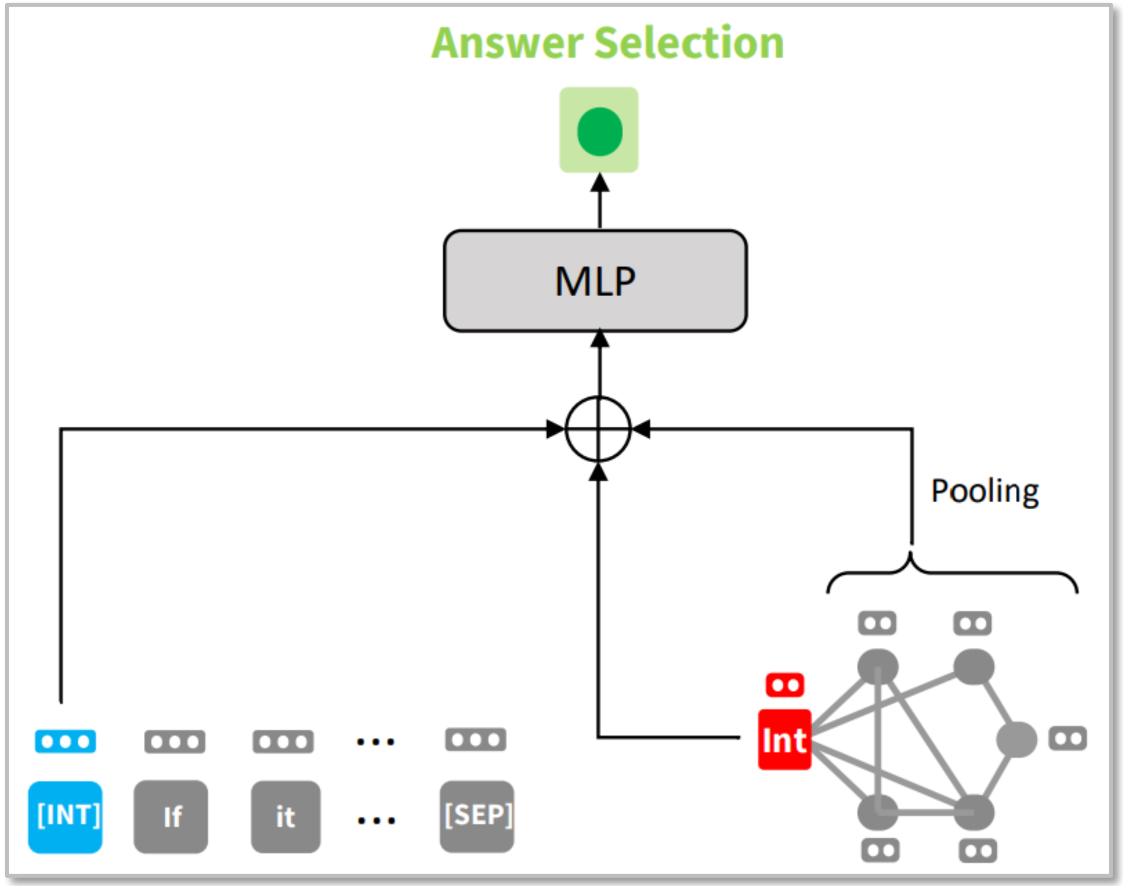
GreaseLM: Deep Bidirectional Cross-Modal Encoder

- Fusing text tokens and KG nodes bidirectionally for multiple layers



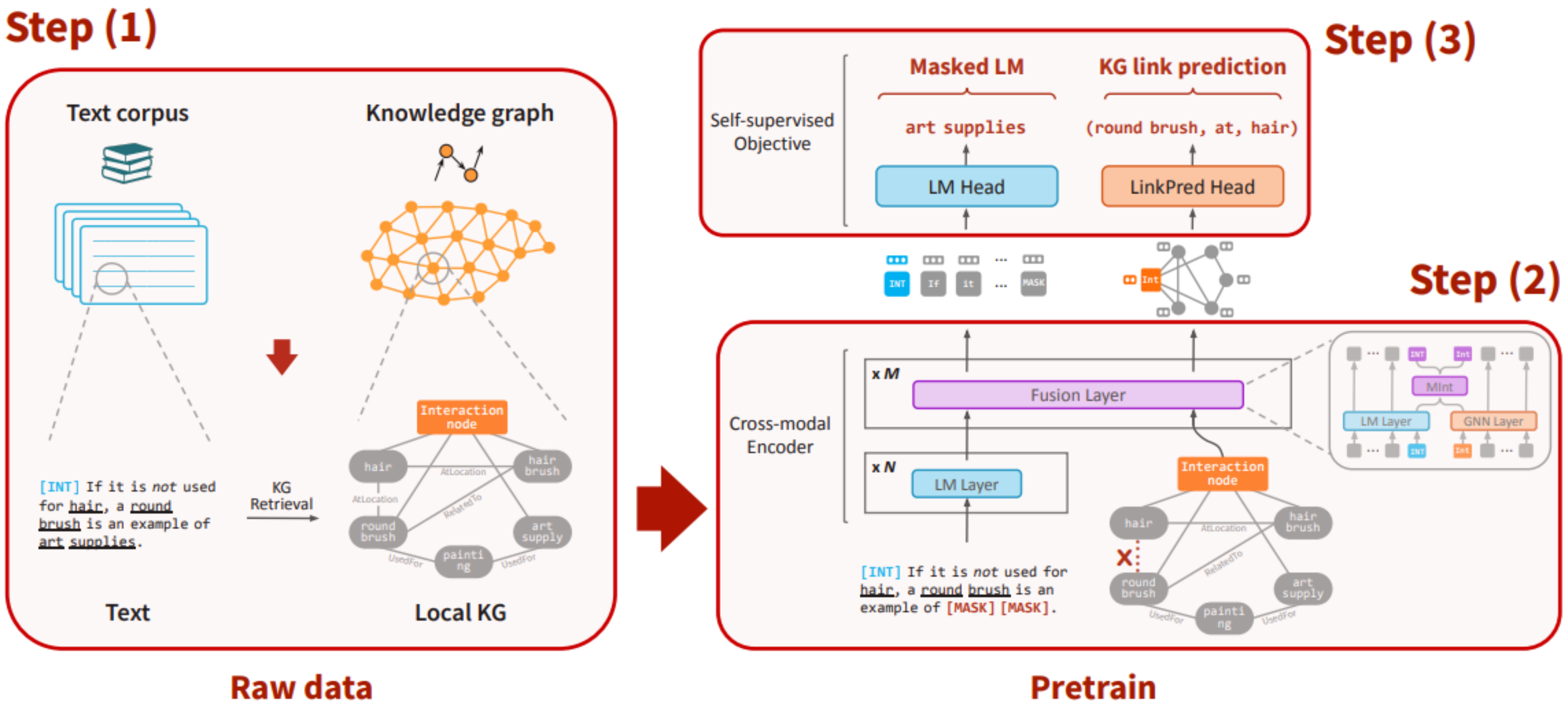
GreaseLM: Multiple-Choice Question Answering

Dataset	Example
CommonsenseQA	A weasel has a thin body and short legs to easier burrow after prey in a what? (A) tree (B) mulberry bush (C) chicken coop (D) viking ship **(E) rabbit warren**



Methods	IHtest-Acc. (%)
RoBERTa-Large (w/o KG)	68.7 (±0.6)
RGCN (Schlichtkrull et al., 2018)	68.4 (±0.7)
GconAttn (Wang et al., 2019)	68.6 (±1.0)
KagNet (Lin et al., 2019)	69.0 (±0.8)
RN (Santoro et al., 2017)	69.1 (±0.2)
MHGRN (Feng et al., 2020)	71.1 (±0.8)
QA-GNN (Yasunaga et al., 2021)	73.4 (±0.9)
GREASELM	**74.2** (±0.4)

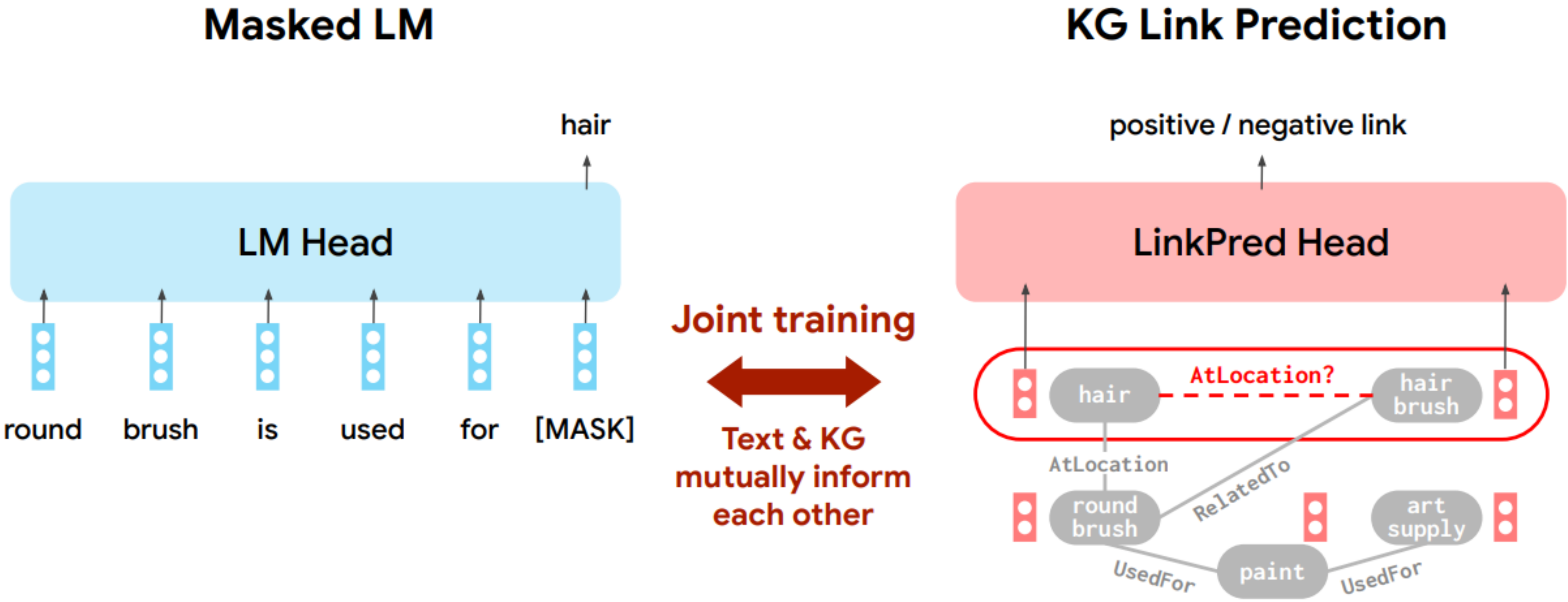
DRAGON: Combining Text & KG for Pre-training



Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C., Liang, P., & Leskovec, J., "Deep Bidirectional Language-Knowledge Graph Pretraining." NeurIPS'22.

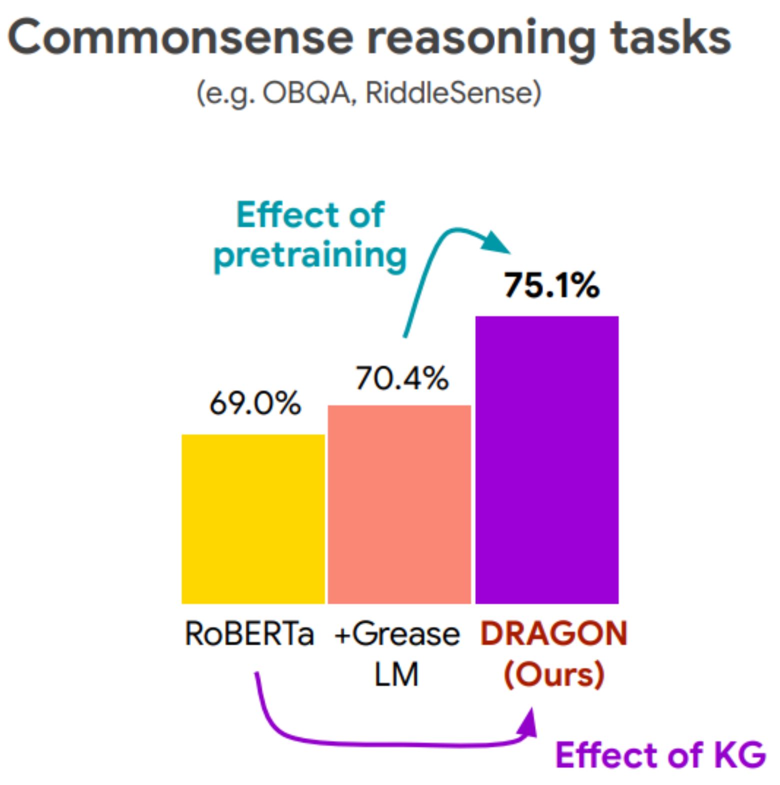
DRAGON: From Multiple-Choice to Self-Supervision

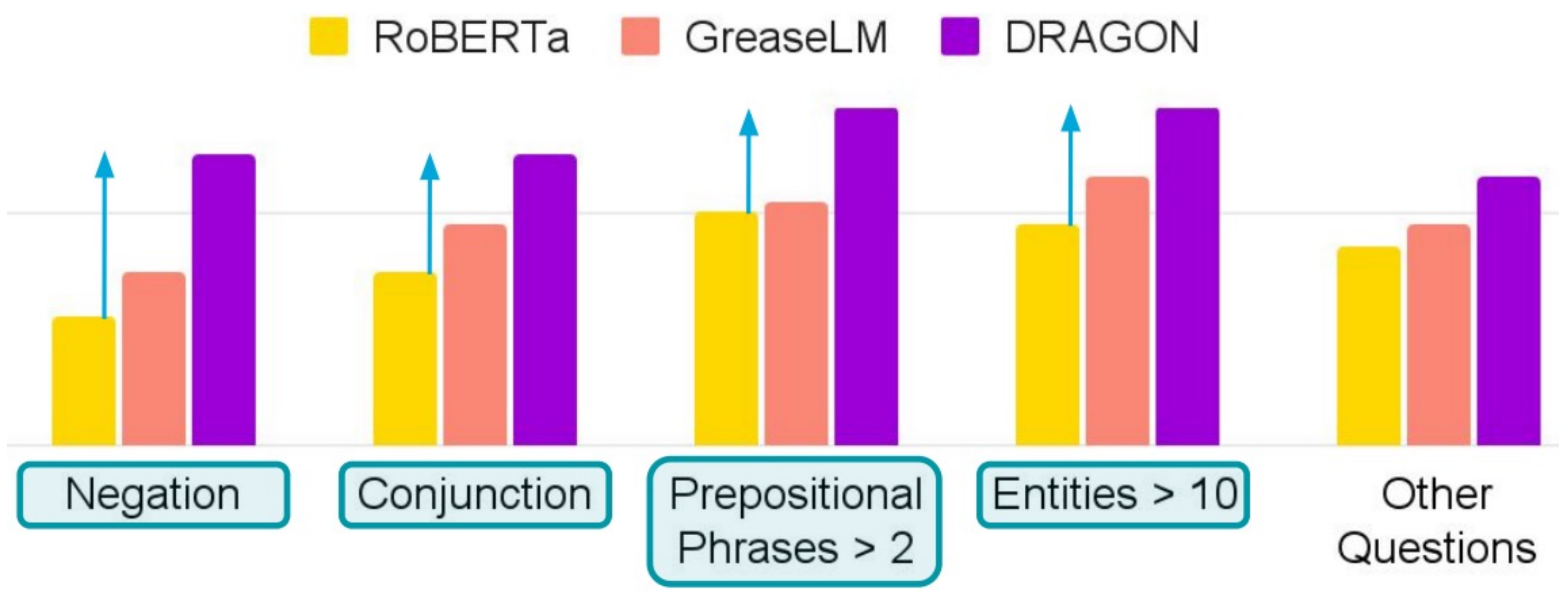
- Pre-training with two self-supervised reasoning tasks



DRAGON: Experimental Results

- Large gains on QA examples involving complex reasoning





Outline

- Structure-enhanced Text Classification
- Structure-enhanced Question Answering
- Structure-enhanced Language Model Pre-training
  - Citation Link
  - Integrating Multiple Types of Structured Information

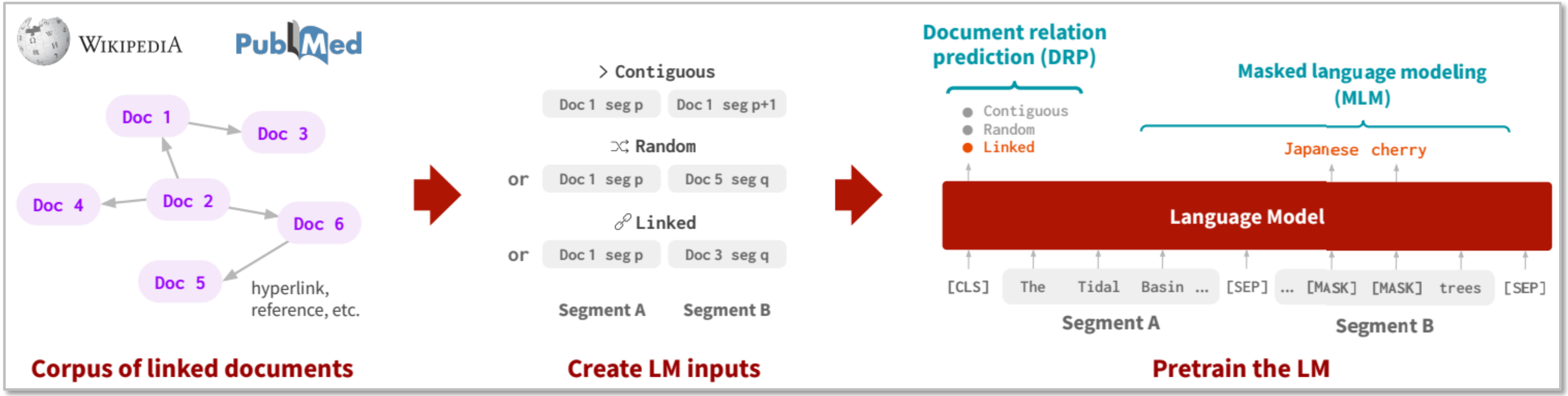
Citation Link Information

- Available in scientific papers, Wikipedia articles, webpages, …
- Benefitting classification, recommendation, question answering, …
- Capturing knowledge and semantics not reflected in the local context within each document

Yasunaga, M., Leskovec, J., & Liang, P., “LinkBERT: Pretraining Language Models with Document Links.” ACL’22.

LinkBERT: A Cross-Encoder Architecture

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



Yasunaga, M., Leskovec, J., & Liang, P., "LinkBERT: Pretraining Language Models with Document Links." ACL'22.

LinkBERT: Experimental Results

- Outperforming BERT in extractive question answering

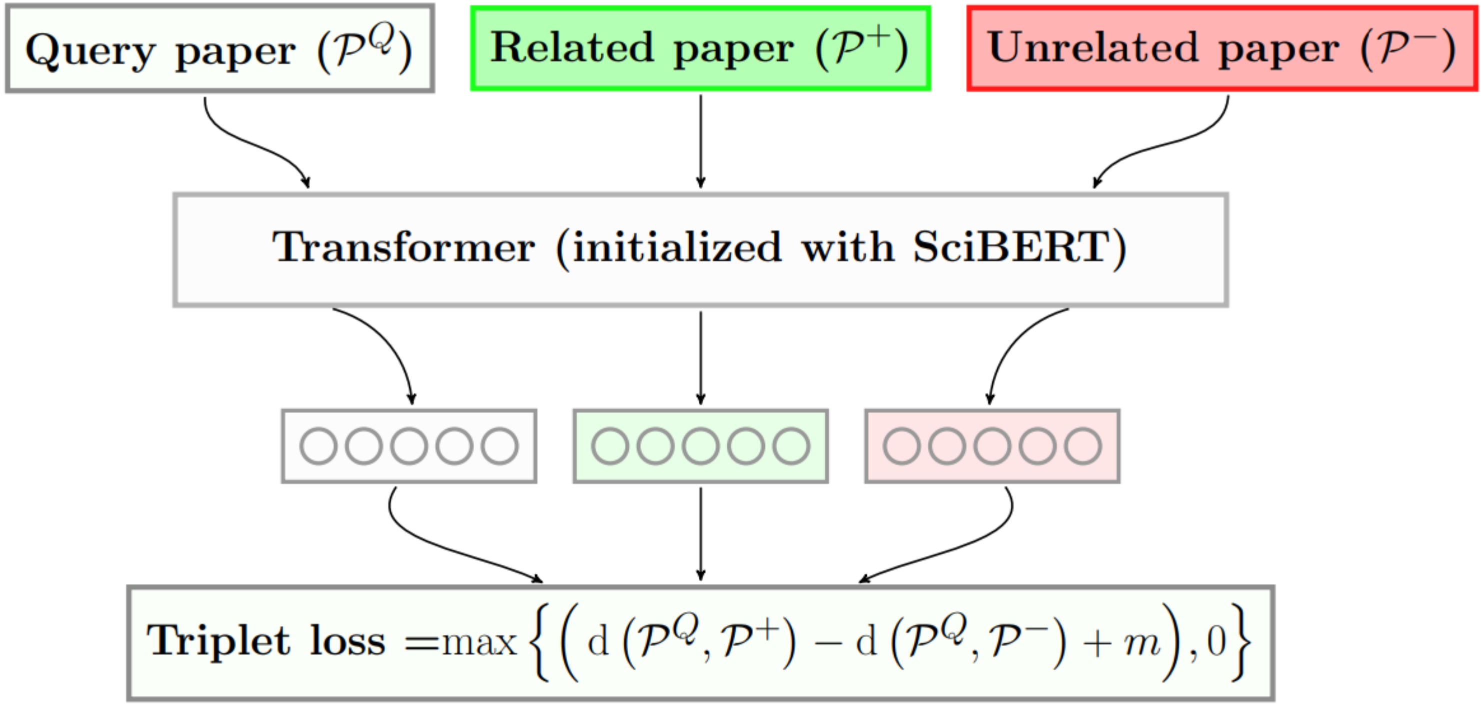
	HotpotQA	TriviaQA	SearchQA	NaturalQ	NewsQA	SQuAD	Avg.
$\text{BERT}_{\texttt{tiny}}$	49.8	43.4	50.2	58.9	41.3	56.6	50.0
$\text{LinkBERT}_{\texttt{tiny}}$	**54.6**	**50.0**	**58.6**	**60.3**	**42.8**	**58.0**	**54.1**
$\text{BERT}_{\texttt{base}}$	76.0	70.3	74.2	76.5	65.7	88.7	75.2
$\text{LinkBERT}_{\texttt{base}}$	**78.2**	**73.9**	**76.8**	**78.3**	**69.3**	**90.1**	**77.8**
$\text{BERT}_{\texttt{large}}$	78.1	73.7	78.3	79.0	70.9	91.1	78.5
$\text{LinkBERT}_{\texttt{large}}$	**80.8**	**78.2**	**80.5**	**81.0**	**72.6**	**92.7**	**81.0**

- Outperforming BERT in natural language understanding tasks (sentiment analysis, NLI, …)

	GLUE score
$\text{BERT}_{\texttt{tiny}}$	64.3
$\text{LinkBERT}_{\texttt{tiny}}$	**64.6**
$\text{BERT}_{\texttt{base}}$	79.2
$\text{LinkBERT}_{\texttt{base}}$	**79.6**
$\text{BERT}_{\texttt{large}}$	80.7
$\text{LinkBERT}_{\texttt{large}}$	**81.1**

SPECTER: A Bi-Encoder Architecture

- Contrastive pre-training via citation prediction
- How to find hard negative samples?
  - IF A cites B
  - AND B cites C
  - BUT A does not cite C
  - THEN C is a hard negative
- Combination of easy and hard negative samples
  - 60% easy + 40% hard

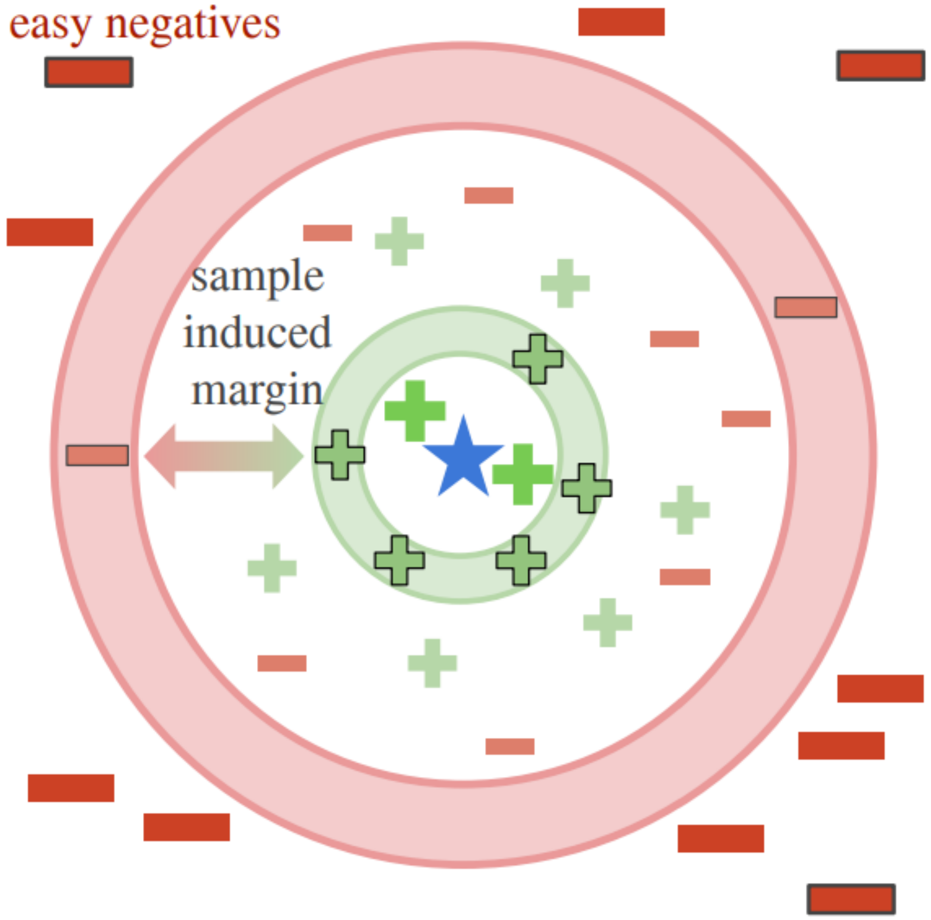


Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D., "SPECTER: Document-level Representation Learning using Citation-informed Transformers." ACL'20.

SciNCL: Improving Hard Negative Sampling

- SPECTER relies on 1 or 2 citation links to obtain positive/negative samples.
- How about a holistic view of the citation graph?

- SciNCL first learns the node embedding of each document based on the citation graph.
  - : query
  - : easy positive (should NOT be used)
  - : hard positive (should be used)
  - confusing area (should NOT be used)
  - : hard negative (should be used)
  - : easy negative



Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., & Rehm, G., "Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings." EMNLP'22.

SPECTER and SciNCL: Experimental Results

- Citation information helps classification, user activity prediction, and recommendation.

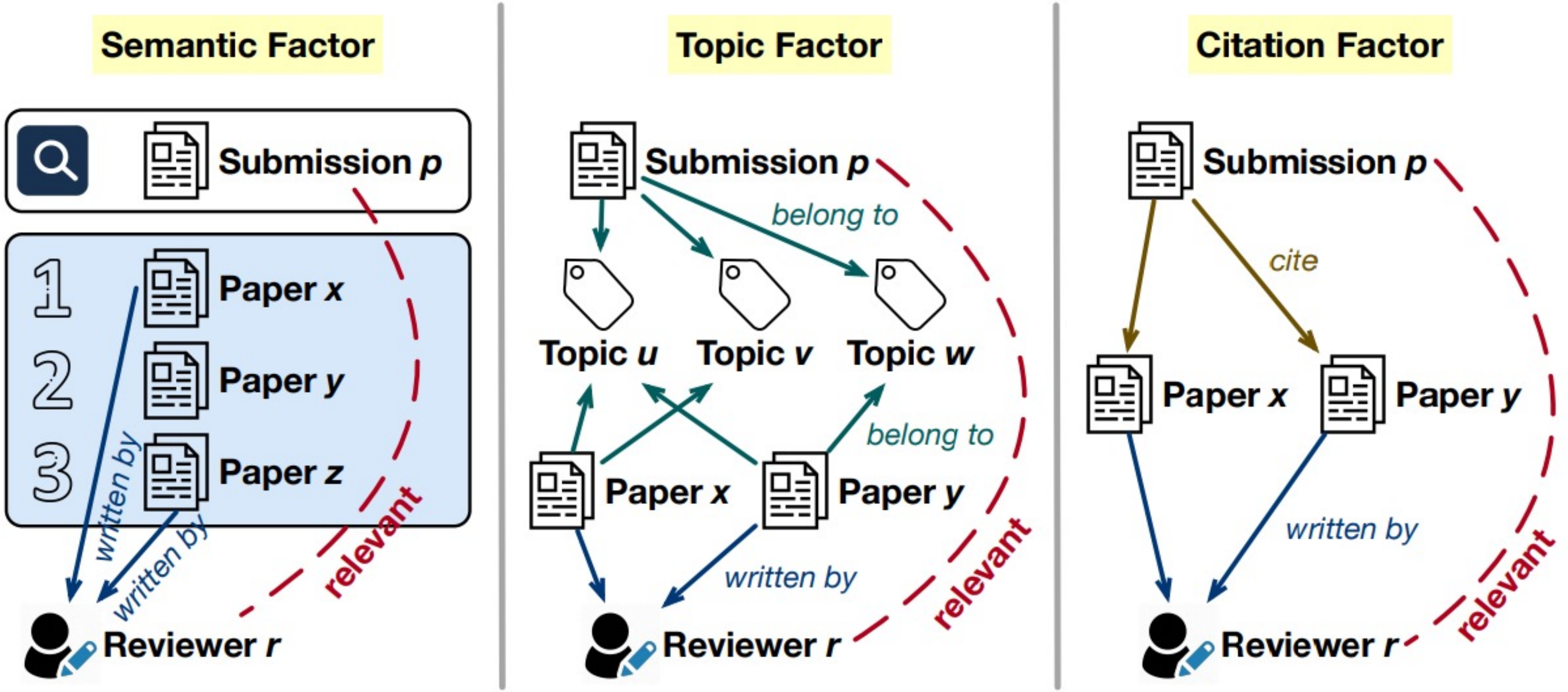
Task →	Classification		User activity prediction				Citation prediction				Recomm.		Avg.
Subtask →	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model ↓ / Metric →	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDCG	P@1	
Oracle SciDocs †	*87.1*	*94.8*	*87.2*	*93.5*	*88.7*	*94.6*	*92.3*	*96.8*	*91.4*	*96.4*	*53.8*	*19.4*	*83.0*
USE (2018)	80.0	83.9	77.2	88.1	76.5	88.1	76.6	89.0	78.3	89.8	53.7	19.6	75.1
Citeomatic* (2018)	67.1	75.7	81.1	90.2	80.5	90.2	86.3	94.1	84.4	92.8	52.5	17.3	76.0
SGC* (2019)	76.8	82.7	77.2	88.0	75.7	87.5	91.6	96.2	84.1	92.5	52.7	18.2	76.9
BERT (2019)	79.9	74.3	59.9	78.3	57.1	76.4	54.3	75.1	57.9	77.3	52.1	18.1	63.4
SciBERT* (2019)	79.7	80.7	50.7	73.1	47.7	71.1	48.3	71.7	49.7	72.6	52.1	17.9	59.6
BioBERT (2019)	77.2	73.0	53.3	74.0	50.6	72.2	45.5	69.0	49.4	71.8	52.0	17.9	58.8
CiteBERT (2021)	78.8	74.8	53.2	73.6	49.9	71.3	45.0	67.9	50.3	72.1	51.6	17.0	58.8
DeCLUTR (2021)	81.2	88.0	63.4	80.6	60.0	78.6	57.2	77.4	62.9	80.9	52.0	17.4	66.6
SPECTER* (2020)	**82.0**	86.4	83.6	91.5	84.5	92.4	88.3	94.9	88.1	94.8	**53.9**	**20.0**	80.0
SciNCL (ours)	81.4	**88.7**	**85.3**	**92.3**	**87.5**	**93.9**	**93.6**	**97.3**	**91.6**	**96.4**	**53.9**	19.3	**81.8**

Outline

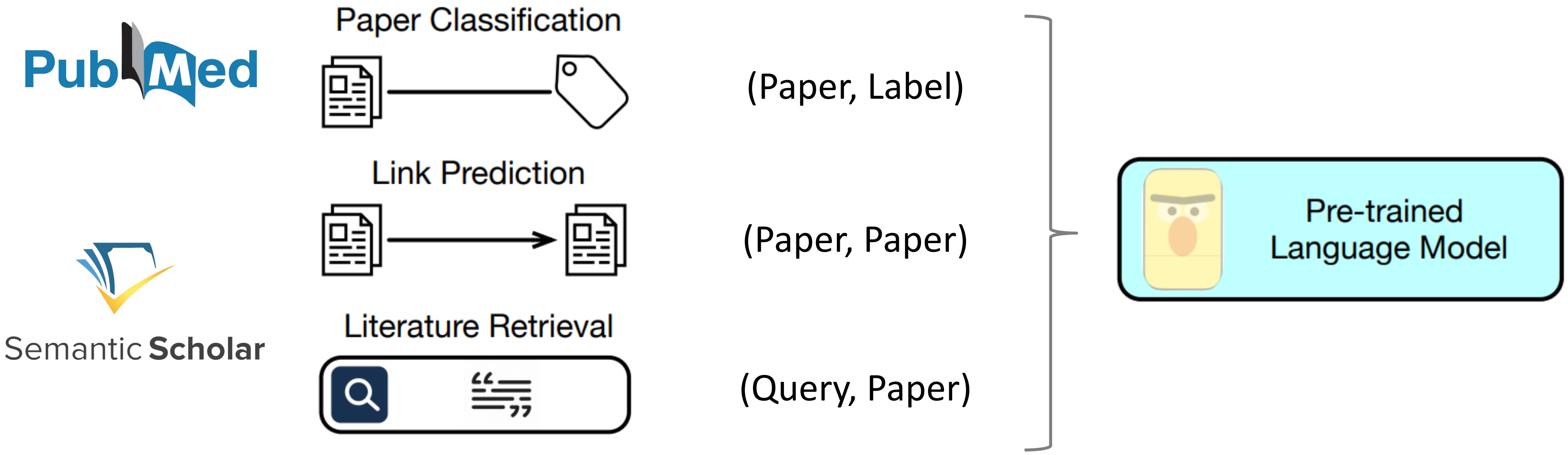
- Structure-enhanced Text Classification
- Structure-enhanced Question Answering
- Structure-enhanced Language Model Pre-training
  - Citation Link
  - Integrating Multiple Types of Structured Information

Multiple Factors when Judging Relevance

- Example: Paper-Reviewer Matching
  - Why is a pair of (Paper, Reviewer) relevant?



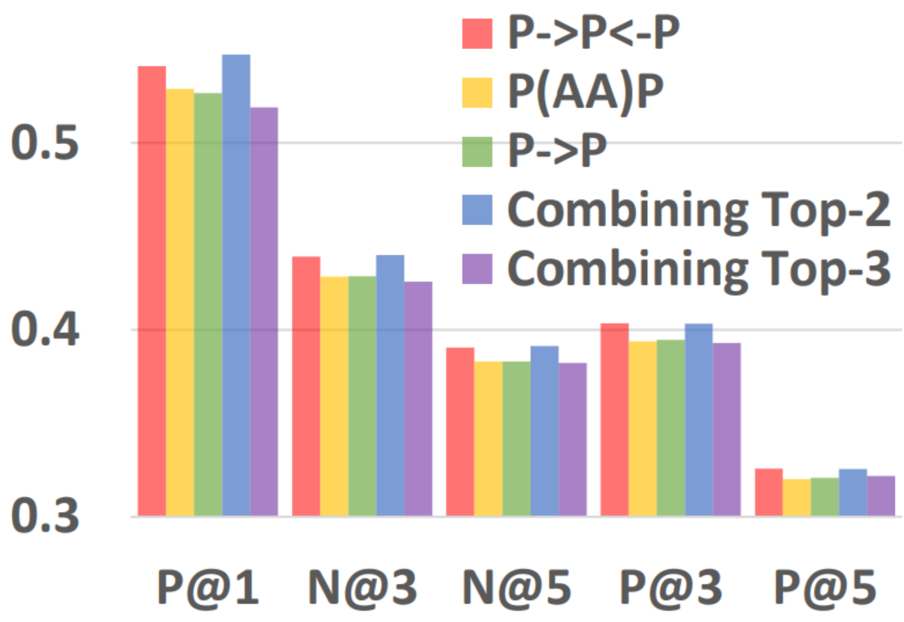
Multiple Types of Available Information

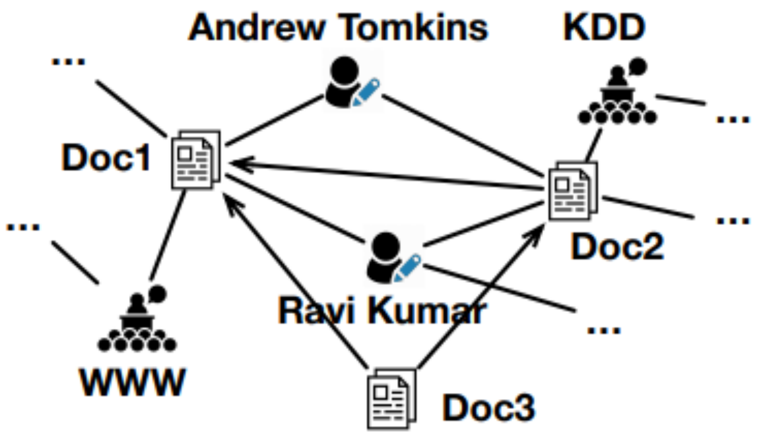


- Directly combining pre-training data from different tasks to train a model?
- Task Interference: The model is confused by different types of "relevance".

An Illustrative Example of Task Interference

- Recall metadata-induced contrastive learning
  - Imagine each meta-path/meta-graph is a "task" (i.e., defines one type of "relevance")
  - Directly merging the relevant (paper, paper) pairs induced by different meta-paths for training?
  - Cannot consistently improve the classification performance!

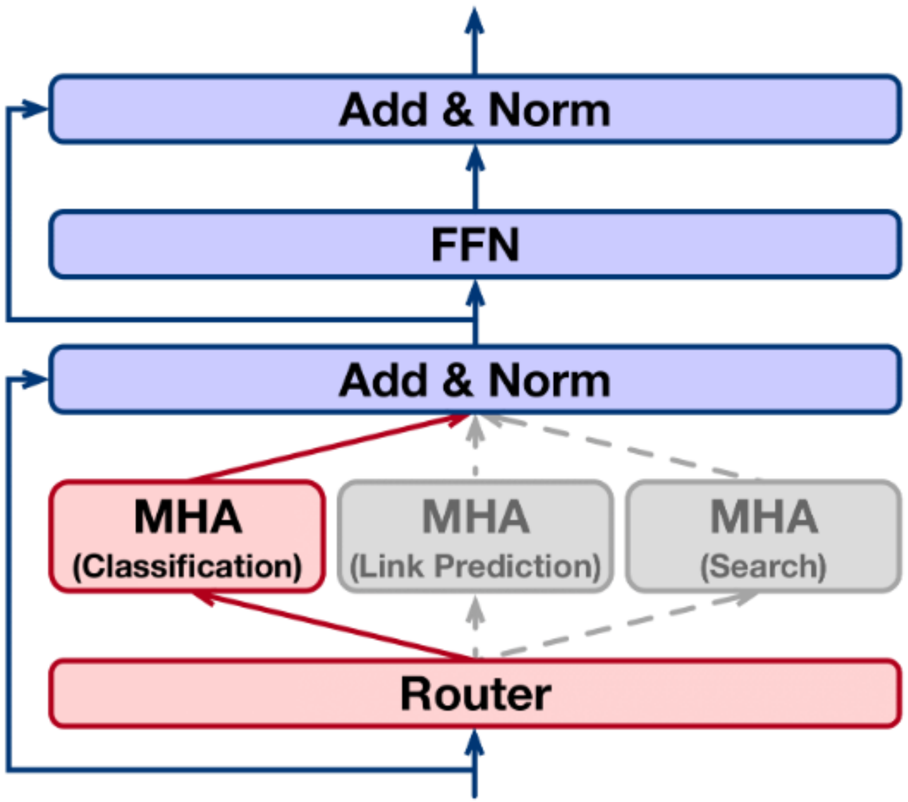




(Doc2, Doc3) are relevant according to P→P←P but irrelevant according to P(AA)P.

Tackling Task Interference: Mixture-of-Experts Transformer (SciMult)

- A typical Transformer layer
  - 1 Multi-Head Attention (MHA) sublayer
  - 1 Feed Forward Network (FFN) sublayer
- A Mixture-of-Experts (MoE) Transformer layer
  - **Multiple** MHA sublayers
  - 1 FFN sublayer
  - (Or 1 MHA & Multiple FFN)
- Specializing some parts of the architecture to be an "expert" of one task
- The model can learn both commonalities and characteristics of different tasks.



Mixture-of-Experts Transformer with Task-Specific **MHA** Sublayers

Zhang, Y., Cheng, H., Shen, Z., Liu, X., Wang, Y.-Y., & Gao, J., "Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding." EMNLP'23 Findings.

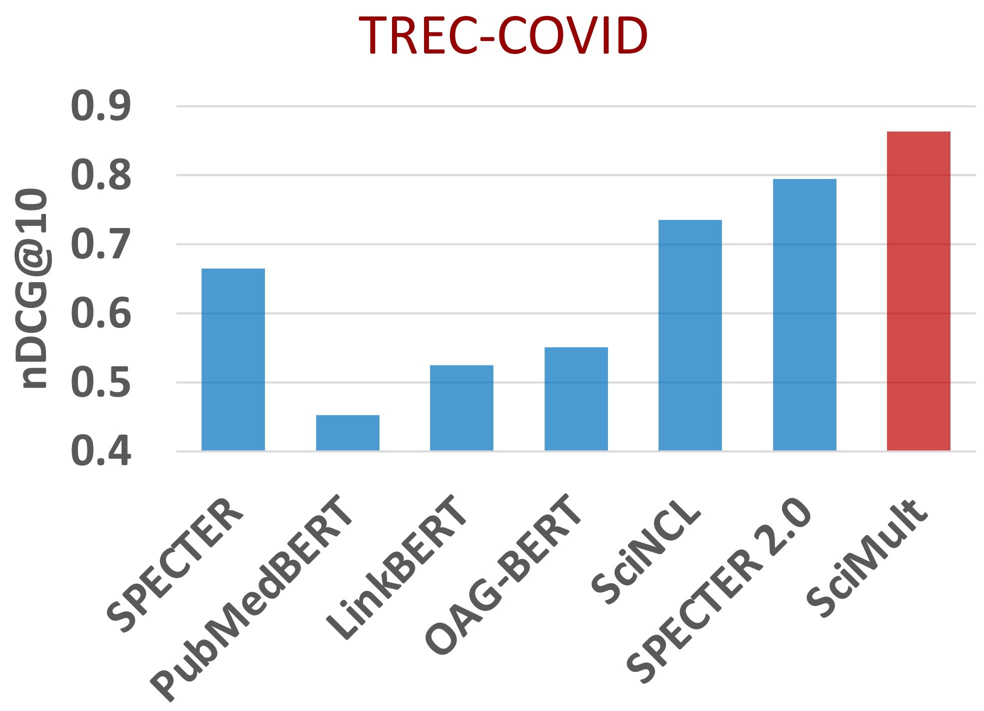
SciMult: Experimental Results

- New SOTA on the PMC-Patients benchmark
- Outperforming previous scientific pre-trained language models in classification, link prediction, literature retrieval, paper recommendation, and claim verification

Patient-to-Article Retrieval (PAR) Leaderboard

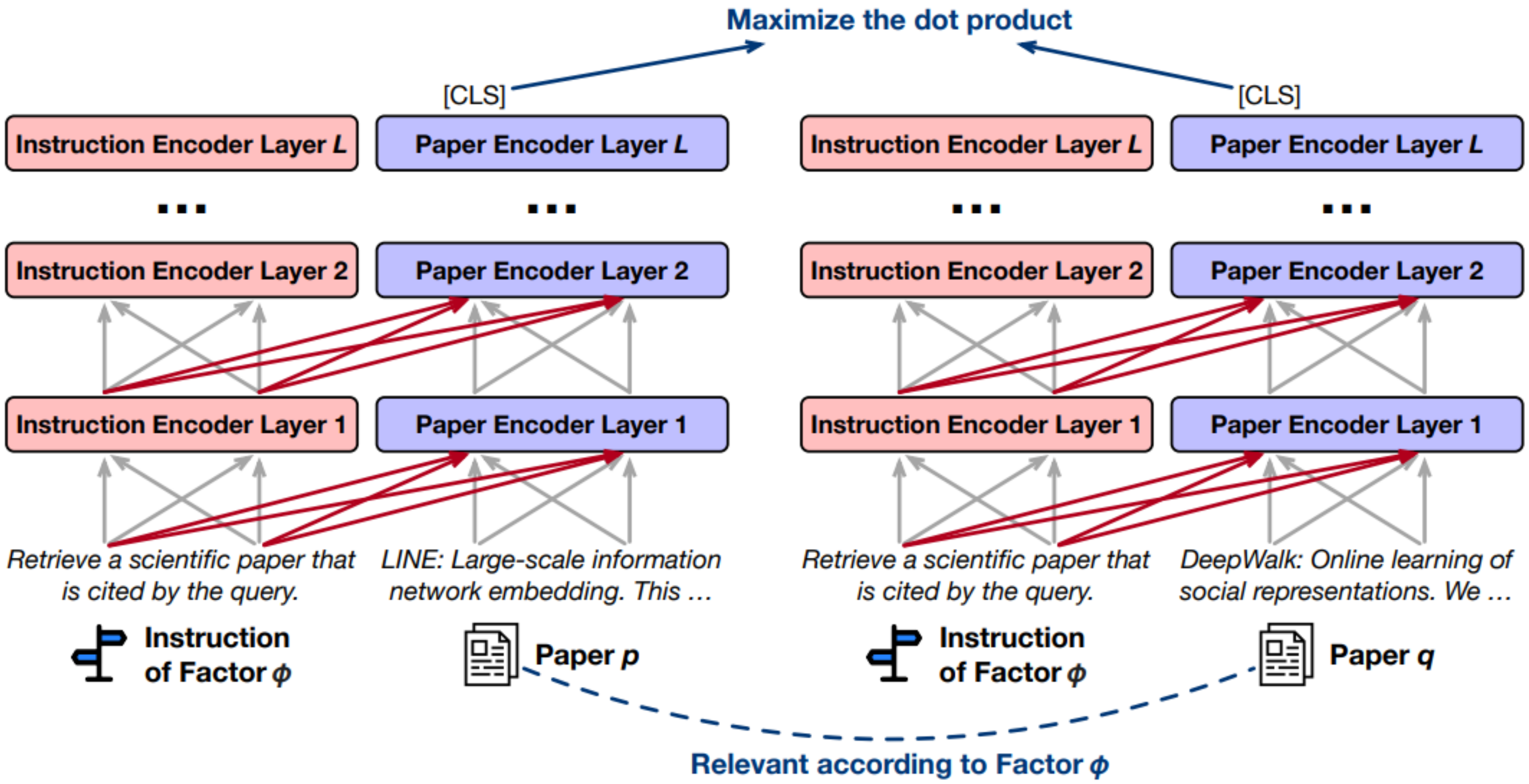
	Model	MRR (%)	P@10 (%)	nDCG@10 (%)	R@1k (%)
1 June 25, 2023	DPR (SciMult-MHAExpert) *UIUC/Microsoft* (Zhang et al. 2023)	29.89	9.35	13.79	53.71
2 Apr 5, 2023	RRF *Tsinghua University* (Zhao et al. 2023)	29.86	8.86	13.36	49.45

https://pmc-patients.github.io/



Tackling Task Interference: Instruction Tuning (UniPR)

- Using a factor-specific instruction to guide the paper encoding process
- The instruction serves as the context of the paper.
- The paper does NOT serve as the context of the instruction.



Zhang, Y., Shen, Y., Chen, X., Jin, B., & Han, J., ""Why Should I Review This Paper?" Unifying Semantic, Topic, and Citation Factors for Paper-Reviewer Matching." arXiv'23.

UniPR: Experimental Results

- Public benchmark datasets
  - Expert C judges whether Reviewer A is qualified to review Paper B.
- Outperforming previous pre-trained scientific language models and the Toronto Paper Matching System (TPMS, used by Microsoft CMT)

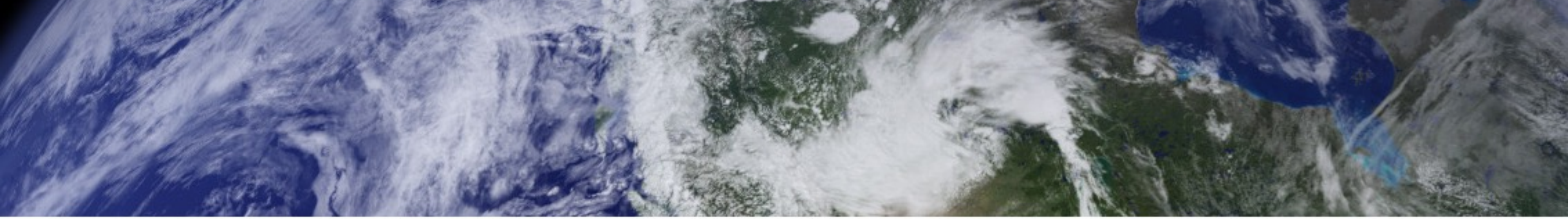
	SciRepEval [48]					SIGIR [22]					KDD				
	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average	Soft P@5	Soft P@10	Hard P@5	Hard P@10	Average
TPMS [8]	62.06	53.74	31.40	24.86	43.02	39.73	38.36	17.81	17.12	28.26	17.01	16.78	6.78	7.24	11.95
SciBERT [6]	59.63	54.39	28.04	24.49	41.64	34.79	34.79	14.79	15.34	24.93	28.51	27.36	12.64	12.70	20.30
SPECTER [10]	65.23	**56.07**	32.34	25.42	44.77	39.73	40.00	16.44	16.71	28.22	34.94	30.52	15.17	**13.28**	23.48
SciNCL [37]	66.92	55.42	34.02	25.33	45.42	**40.55**	39.45	17.81	17.40	28.80	36.21	30.86	15.06	12.70	23.71
COCO-DR [61]	65.05	55.14	31.78	24.67	44.16	40.00	**40.55**	16.71	**17.53**	28.70	35.06	29.89	13.68	12.13	22.69
SPECTER 2.0 CLF [48]	64.49	55.23	31.59	24.49	43.95	39.45	38.63	16.16	16.30	27.64	34.37	30.63	14.48	12.64	23.03
SPECTER 2.0 PRX [48]	66.36	55.61	34.21	**25.61**	45.45	40.00	38.90	19.18	16.85	28.73	37.13	**31.03**	15.86	13.05	24.27
UniPR-NoInstruction	66.73	55.61	34.58	25.42	45.59	39.18	38.77	16.99	15.75	27.67	36.67	30.86	15.75	12.99	24.07
UniPR	**69.16**	55.89	**34.95**	25.42	**46.36**	39.73	40.14	**19.73**	16.99	**29.15**	**37.47**	30.75	**15.98**	**13.28**	**24.37**

References

- Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D., “SPECTER: Document-level Representation Learning using Citation-informed Transformers.” ACL’20.
- Gopal, S., & Yang, Y., “Recursive regularization for large-scale classification with hierarchical and graphical dependencies.” KDD’13.
- Liu, X., Yin, D., Zheng, J., Zhang, X., Zhang, P., Yang, H., Dong, Y., & Tang, J. "OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services", KDD’22.
- Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Zhang, C., & Han, J. "Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding", KDD’20.
- Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., & Rehm, G., “Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings.” EMNLP’22.
- Shen, J., Qiu, W., Meng, Y., Shang, J., Ren, X., & Han, J., “TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names”, NAACL’21.
- Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C., Liang, P., & Leskovec, J., “Deep Bidirectional Language-Knowledge Graph Pretraining.” NeurIPS’22.
- Yasunaga, M., Leskovec, J., & Liang, P., “LinkBERT: Pretraining Language Models with Document Links.” ACL’22.

References

- Zhang, X., Bosselut, A., Yasunaga, M., Ren, H., Liang, P., Leskovec, J., & Manning, C., “GreaseLM: Graph REASoning Enhanced Language Models for Question Answering.” ICLR’22.
- Zhang, Y., Cheng, H., Shen, Z., Liu, X., Wang, Y.-Y., & Gao, J., “Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding.” EMNLP’23 Findings.
- Zhang, Y., Shen, Y., Chen, X., Jin, B., & Han, J., ““Why Should I Review This Paper?” Unifying Semantic, Topic, and Citation Factors for Paper-Reviewer Matching.” arXiv’23.
- Zhang, Y., Shen, Z., Dong, Y., Wang, K. & Han, J. "MATCH: Metadata-Aware Text Classification in A Large Hierarchy", WWW’21.
- Zhang, Y., Shen, Z., Wu, C., Xie, B., Wang, Y., Wang, K. & Han, J. "Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification", WWW’22.



Q&A