CSCE 689 - Special Topics in NLP for Science

Lecture 4: Citation Prediction

Yu Zhang

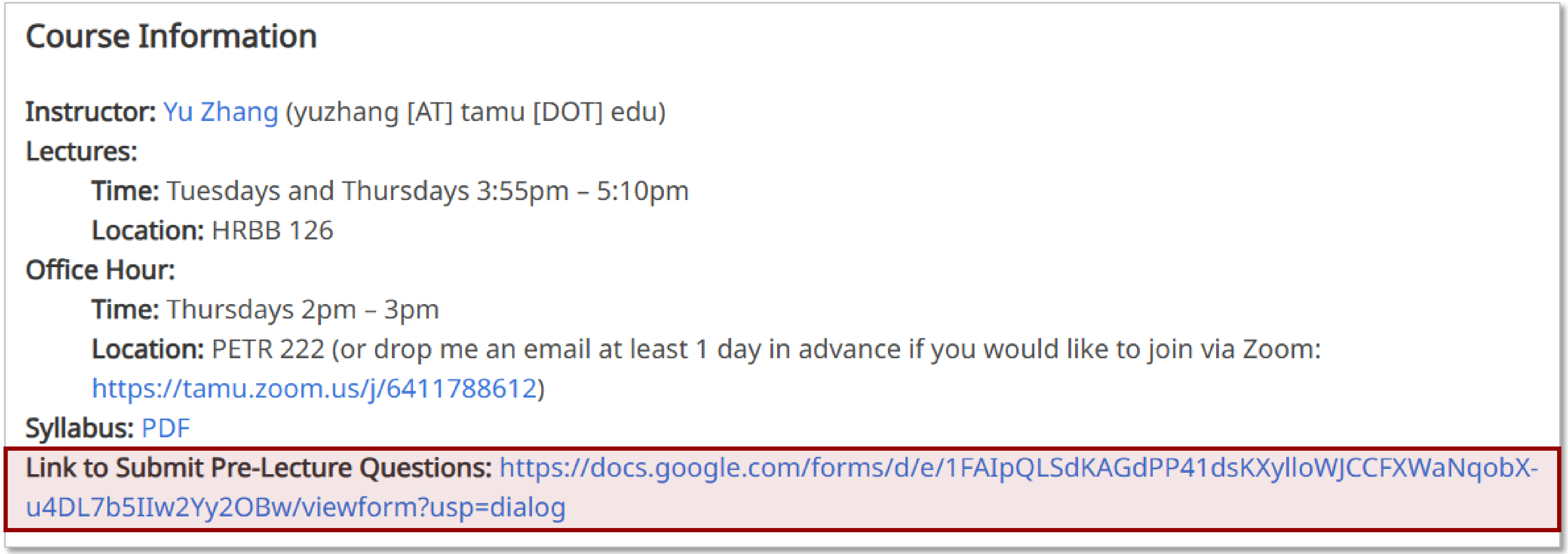
yuzhang@tamu.edu

January 28, 2025

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

Submit Pre-Lecture Questions via Google Form

- https://docs.google.com/forms/d/e/1FAIpQLSdKAGdPP41dsKXylloWJCCFXWaNqobX-u4DL7b5IIw2Yy2OBw/viewform?usp=dialog
- Please submit questions for student lectures and guest lectures only



Course Information

Instructor: Yu Zhang (yuzhang [AT] tamu [DOT] edu)

Lectures:

Time: Tuesdays and Thursdays 3:55pm – 5:10pm

Location: HRBB 126

Office Hour:

Time: Thursdays 2pm – 3pm

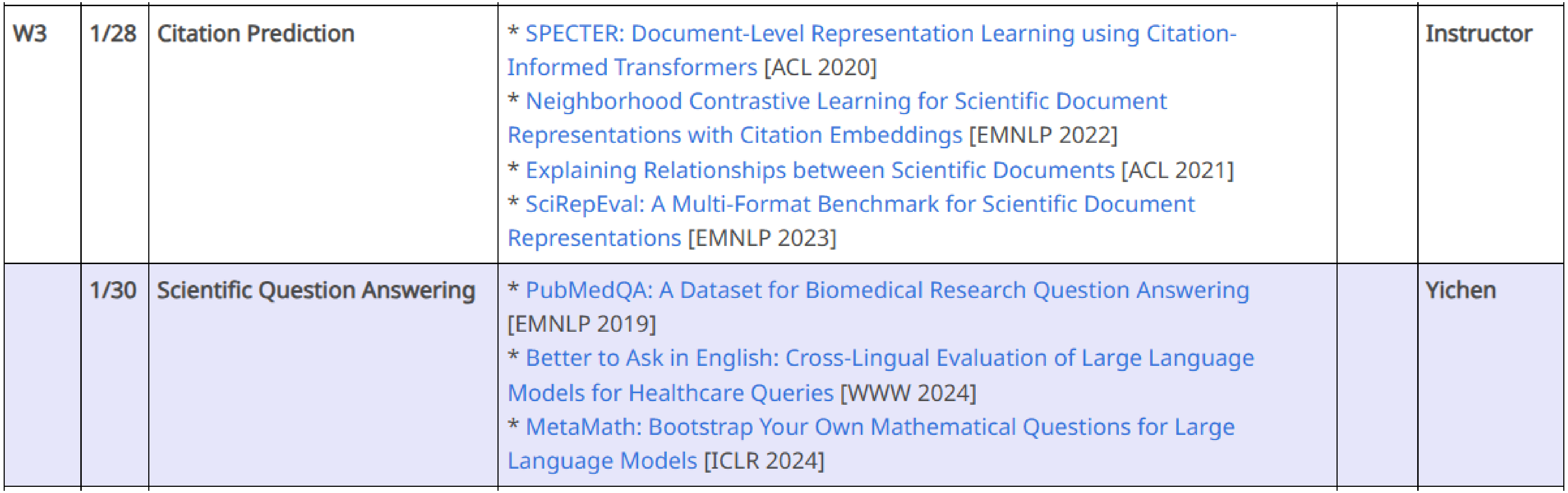
Location: PETR 222 (or drop me an email at least 1 day in advance if you would like to join via Zoom: https://tamu.zoom.us/j/6411788612)

Syllabus: PDF

Link to Submit Pre-Lecture Questions: https://docs.google.com/forms/d/e/1FAIpQLSdKAGdPP41dsKXylloWJCCFXWaNqobX-u4DL7b5IIw2Yy2OBw/viewform?usp=dialog

Submit Pre-Lecture Questions via Google Form

- The first student lecture will be given by Yichen this Thursday.
- If you want to submit a question for Yichen's lecture, the deadline is 11:59pm this Wednesday.
- We will have 10 student lectures + 3 guest lectures, and you only need to submit 5 questions.



W3	1/28	Citation Prediction	* SPECTER: Document-Level Representation Learning using Citation-Informed Transformers [ACL 2020] * Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings [EMNLP 2022] * Explaining Relationships between Scientific Documents [ACL 2021] * SciRepEval: A Multi-Format Benchmark for Scientific Document Representations [EMNLP 2023]		Instructor
	1/30	Scientific Question Answering	* PubMedQA: A Dataset for Biomedical Research Question Answering [EMNLP 2019] * Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries [WWW 2024] * MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models [ICLR 2024]		Yichen

Scientific Papers

- In previous lectures, we mainly utilize the text information (e.g., title, abstract, and full text) of scientific papers to train LLMs.

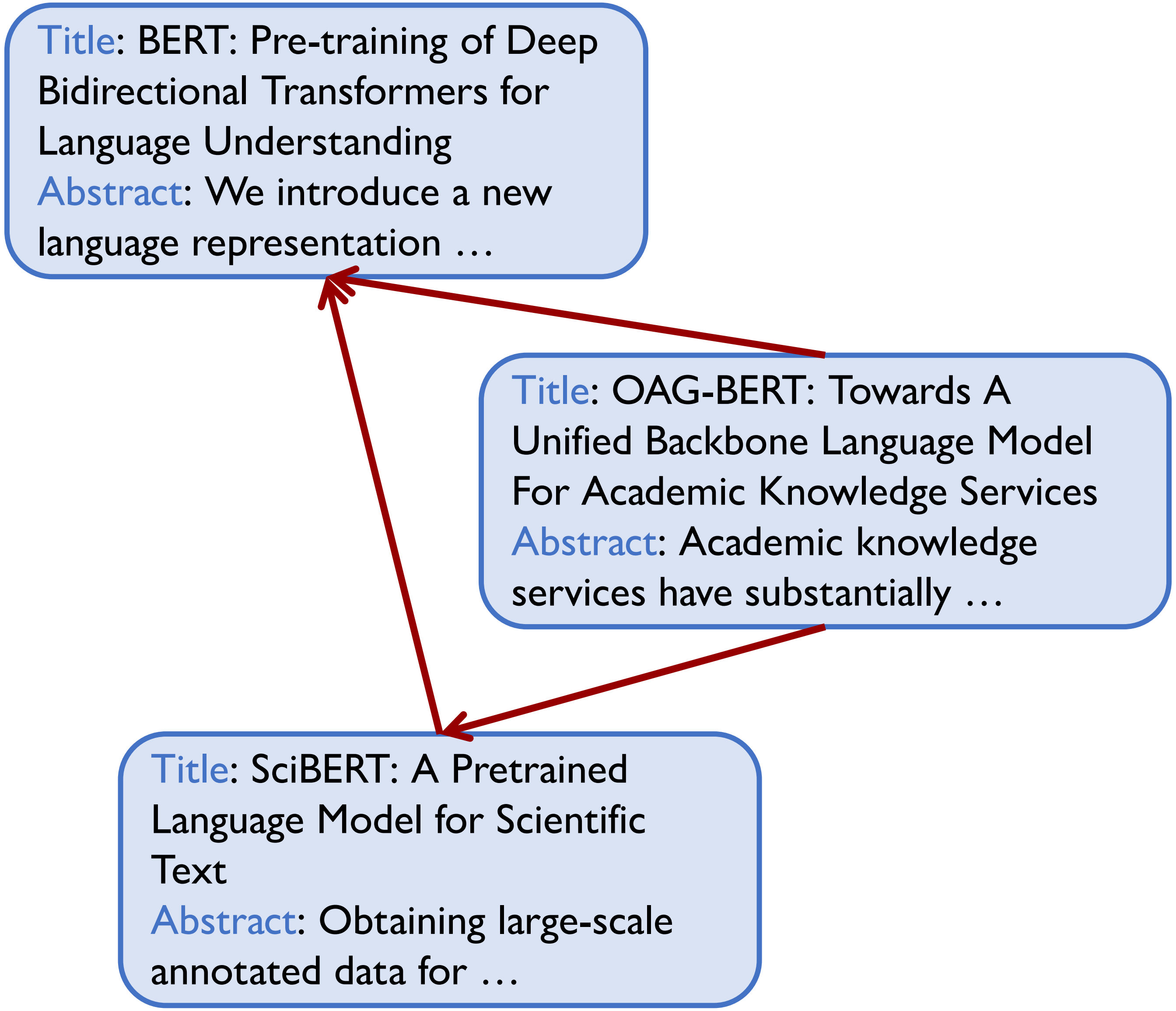
Title: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Abstract: We introduce a new language representation …

Title: OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services
Abstract: Academic knowledge services have substantially …

Title: SciBERT: A Pretrained Language Model for Scientific Text
Abstract: Obtaining large-scale annotated data for …

Scientific Papers

- In previous lectures, we mainly utilize the text information (e.g., title, abstract, and full text) of scientific papers to train LLMs.
- Scientific papers are not plain text sequences. They are associated with:
  - Citation(s) → Today
  - Author(s) → 3/18 & 3/25
  - Venue → 3/18 & 3/25
  - …



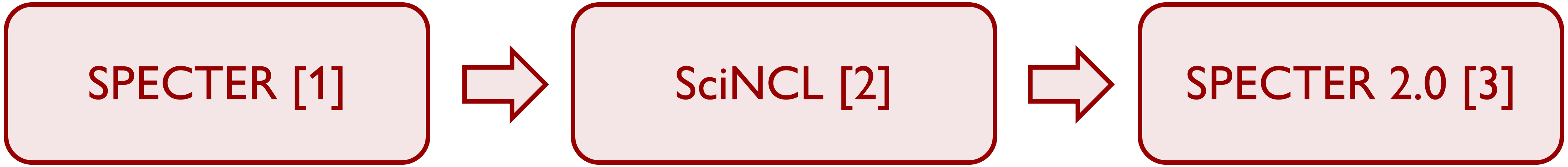
Title: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Abstract: We introduce a new language representation …

Title: OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services
Abstract: Academic knowledge services have substantially …

Title: SciBERT: A Pretrained Language Model for Scientific Text
Abstract: Obtaining large-scale annotated data for …

Two Questions Related to Citations

- Question 1: How to train an LLM to perform citation prediction?
- Question 2: Can citation information help an LLM with other tasks?



SPECTER [1] → SciNCL [2] → SPECTER 2.0 [3]

[1] *SPECTER: Document-Level Representation Learning using Citation-Informed Transformers*. ACL 2020.

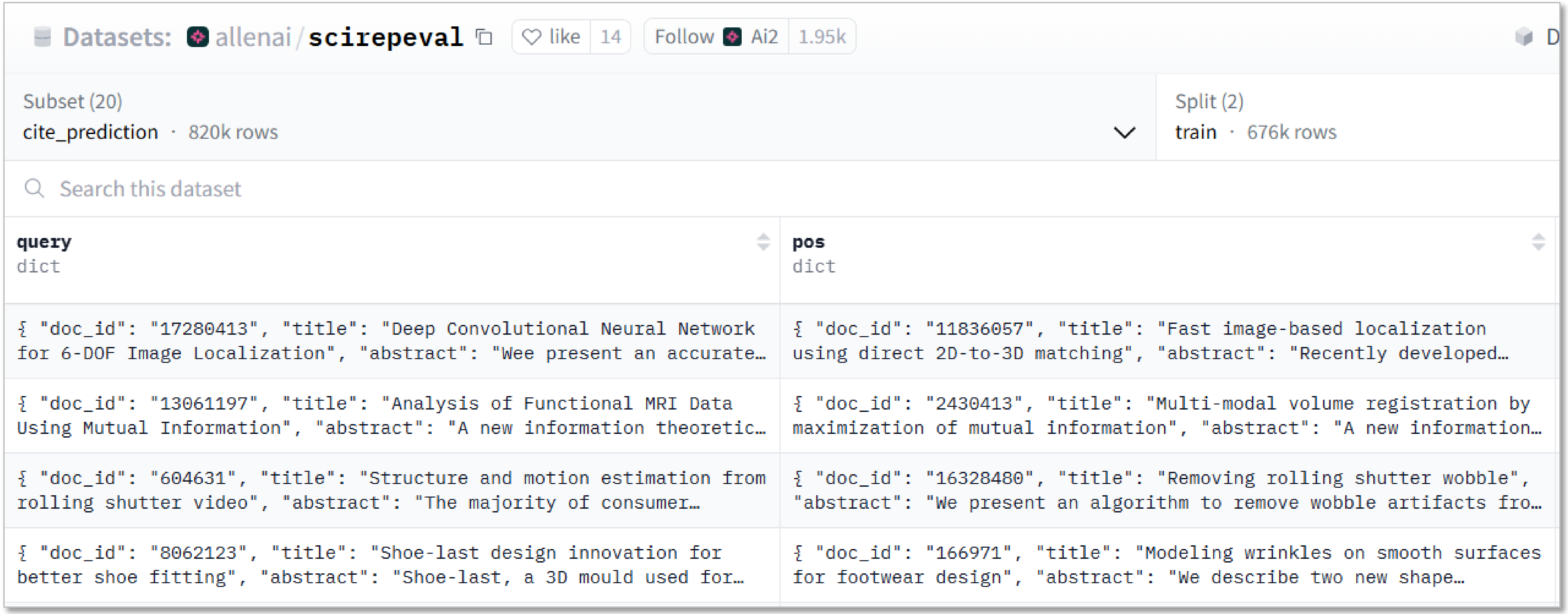
[2] *Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings*. EMNLP 2022.

[3] *SciRepEval: A Multi-Format Benchmark for Scientific Document Representations*. EMNLP 2023.

Q1: How to train an LLM to perform citation prediction?

- Step 1: Collect a large number of papers with citation information.

https://huggingface.co/datasets/allenai/scirepeval/viewer/cite_prediction



Subset (20)
cite_prediction · 820k rows

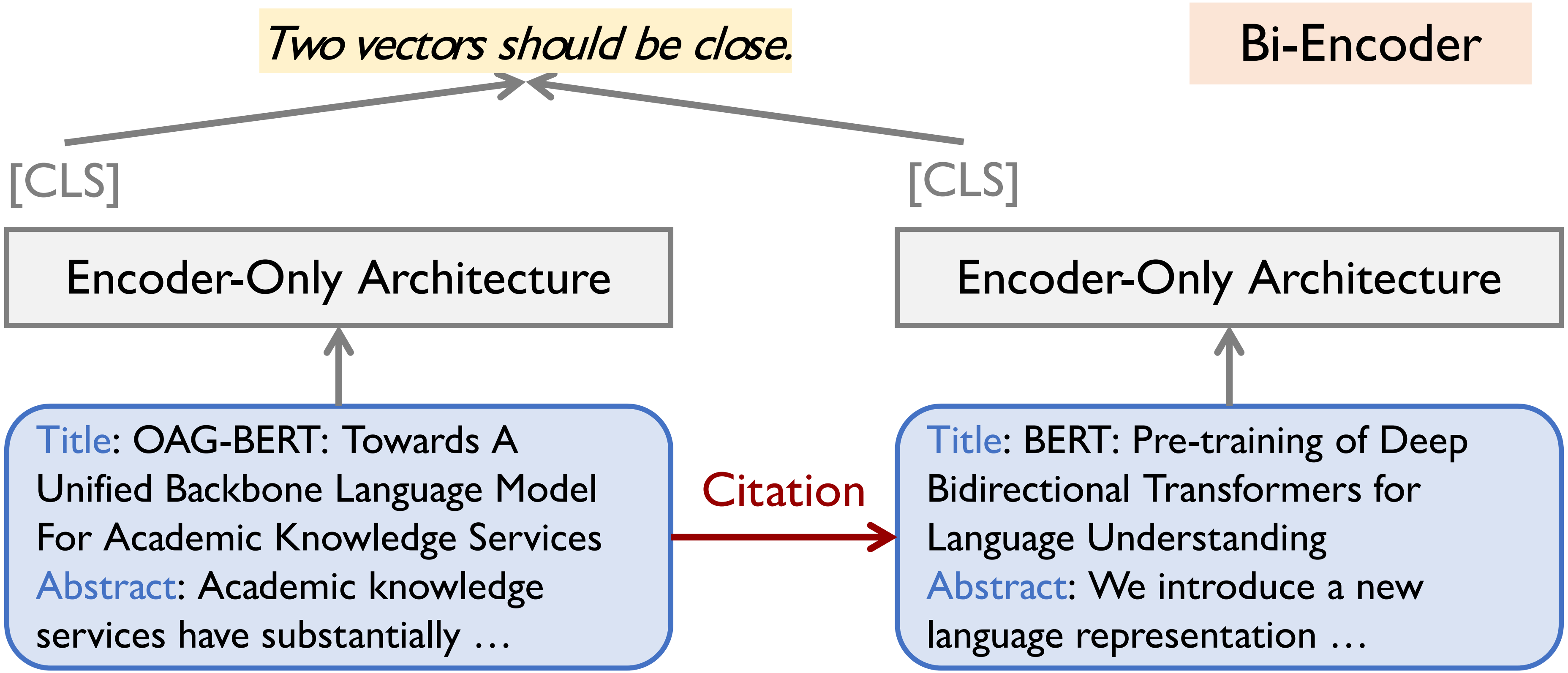
Split (2)
train · 676k rows

Search this dataset

query dict	pos dict
{ "doc_id": "17280413", "title": "Deep Convolutional Neural Network for 6-DOF Image Localization", "abstract": "Wee present an accurate…	{ "doc_id": "11836057", "title": "Fast image-based localization using direct 2D-to-3D matching", "abstract": "Recently developed…
{ "doc_id": "13061197", "title": "Analysis of Functional MRI Data Using Mutual Information", "abstract": "A new information theoretic…	{ "doc_id": "2430413", "title": "Multi-modal volume registration by maximization of mutual information", "abstract": "A new information…
{ "doc_id": "604631", "title": "Structure and motion estimation from rolling shutter video", "abstract": "The majority of consumer…	{ "doc_id": "16328480", "title": "Removing rolling shutter wobble", "abstract": "We present an algorithm to remove wobble artifacts fro…
{ "doc_id": "8062123", "title": "Shoe-last design innovation for better shoe fitting", "abstract": "Shoe-last, a 3D mould used for…	{ "doc_id": "166971", "title": "Modeling wrinkles on smooth surfaces for footwear design", "abstract": "We describe two new shape…

Q1: How to train an LLM to perform citation prediction?

- Step 1: Collect a large number of papers with citation information.
- Step 2: Train an LLM with such citation information.



Two vectors should be close.

Bi-Encoder

[CLS]

Encoder-Only Architecture

Title: OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services
Abstract: Academic knowledge services have substantially …

Citation

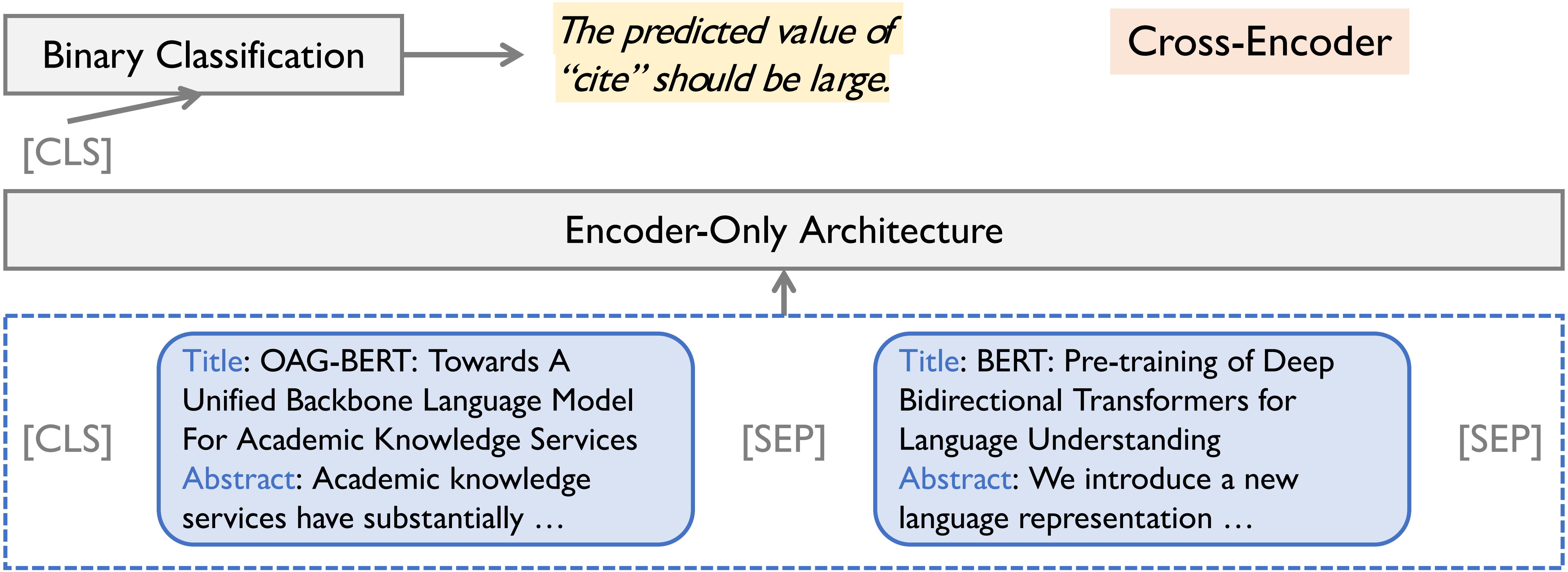
[CLS]

Encoder-Only Architecture

Title: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Abstract: We introduce a new language representation …

Q1: How to train an LLM to perform citation prediction?

- Step 1: Collect a large number of papers with citation information.
- Step 2: Train an LLM with such citation information.



Binary Classification

The predicted value of "cite" should be large.

Cross-Encoder

[CLS]

Encoder-Only Architecture

[CLS]

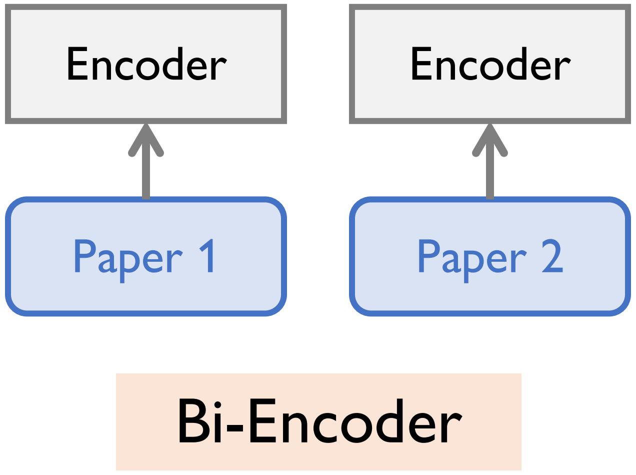
Title: OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services
Abstract: Academic knowledge services have substantially …

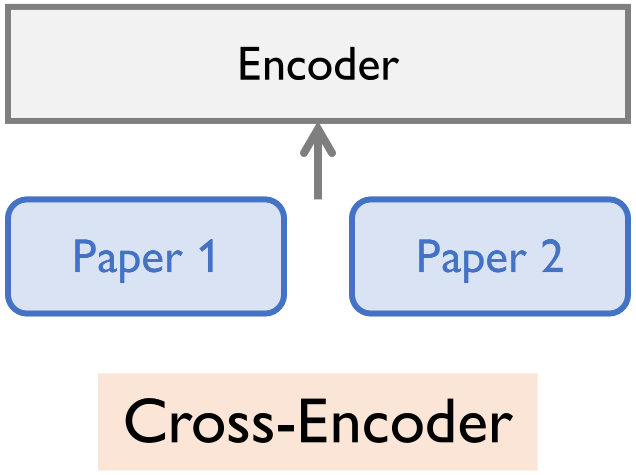
[SEP]

Title: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Abstract: We introduce a new language representation …

[SEP]

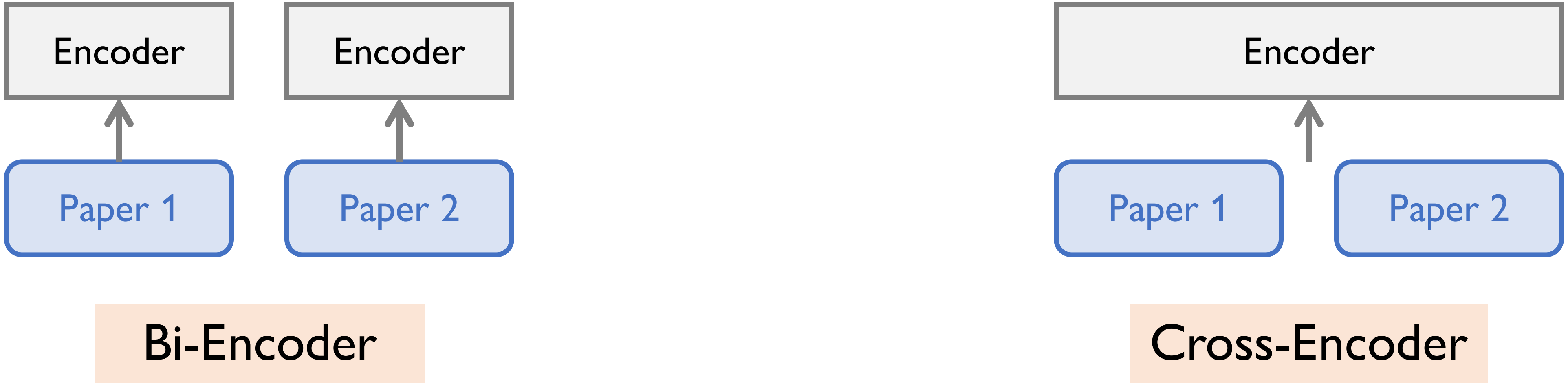
Bi-Encoder vs. Cross-Encoder





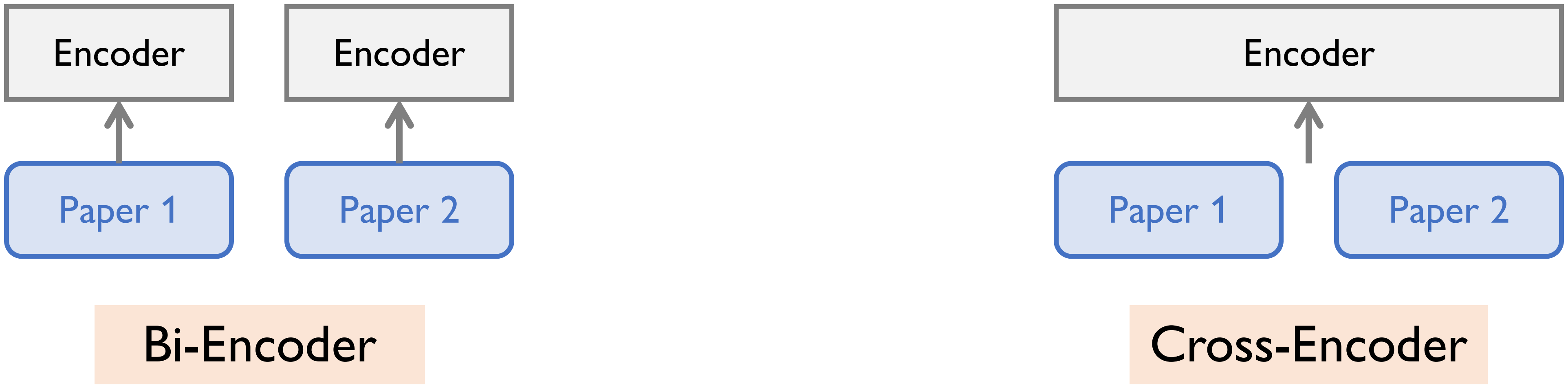
- Advantages of Cross-Encoder
  - The idea is similar to the next sentence prediction task for pre-training BERT/SciBERT. If you start training your model from BERT/SciBERT, the model has had some citation prediction abilities at the beginning.
  - Two papers can serve as context of each other, so that the model can learn a better contextualized representation of each token in the input sequence.

Bi-Encoder vs. Cross-Encoder



- Advantages of Bi-Encoder
  - More text information can be fed into the encoder.
    - Assume one encoder can take at most N tokens. Bi-Encoder truncates each paper at its N-th token. Cross-Encoder truncates each paper text at its 0.5N-th token.

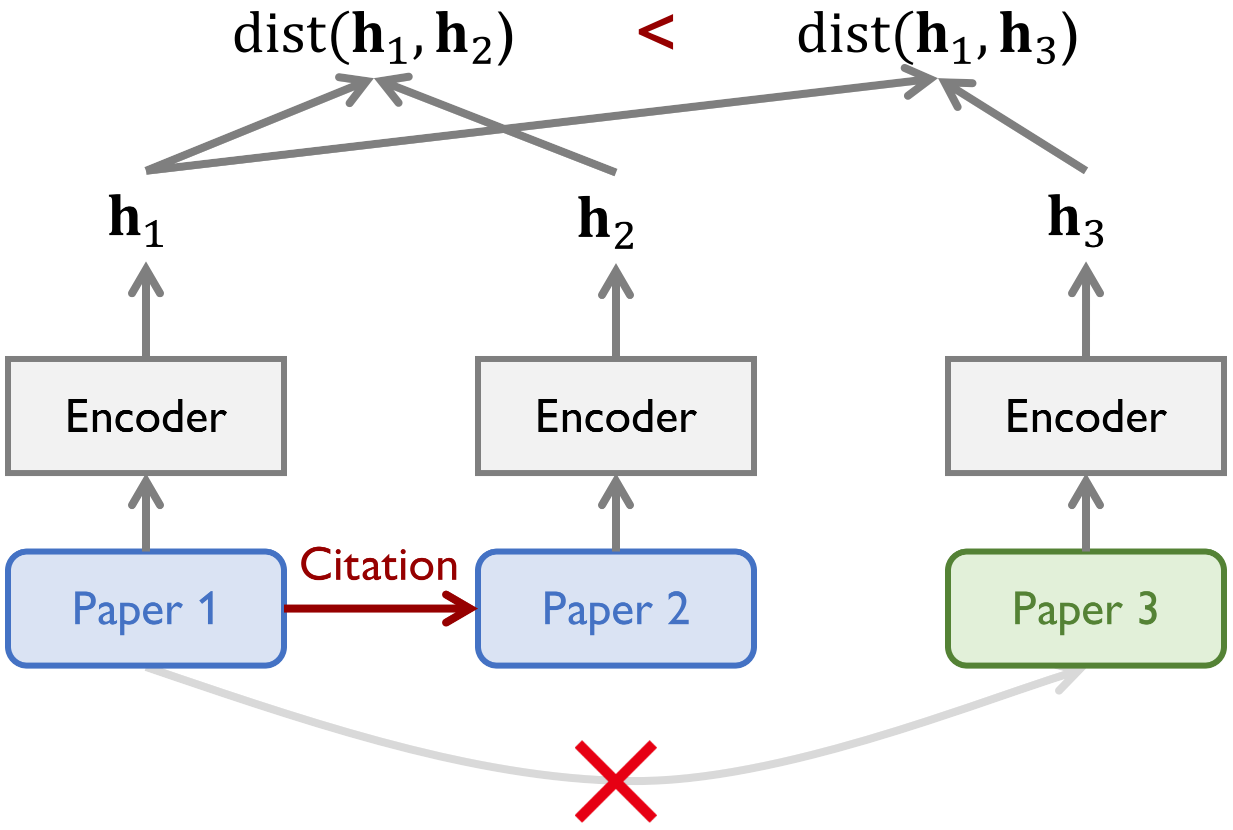
Bi-Encoder vs. Cross-Encoder



- Advantages of Bi-Encoder
  - Bi-Encoder is much more efficient during the inference time.
    - Suppose you have 1,000 papers. How many times do you need to call the trained encoder to make pair-wise predictions?
    - Bi-Encoder: 1,000
    - Cross-Encoder: 1,000 × 1,000 = 1,000,000

Contrastive Learning

- SPECTER, SciNCL, and SPECTER 2.0 all use the Bi-Encoder architecture.



$\text{dist}(\mathbf{h}_1, \mathbf{h}_2)$ < $\text{dist}(\mathbf{h}_1, \mathbf{h}_3)$

$\mathbf{h}_1$ $\mathbf{h}_2$ $\mathbf{h}_3$

Encoder Encoder Encoder

Paper 1 Citation Paper 2 Paper 3

Loss Function: minimize

$$\max\{\text{dist}(\mathbf{h}_1, \mathbf{h}_2) - \text{dist}(\mathbf{h}_1, \mathbf{h}_3) + m, 0\}$$

Other Possible Choices: maximize

$$\frac{\exp(\cos(\mathbf{h}_1, \mathbf{h}_2))}{\exp(\cos(\mathbf{h}_1, \mathbf{h}_2)) + \exp(\cos(\mathbf{h}_1, \mathbf{h}_3))}$$

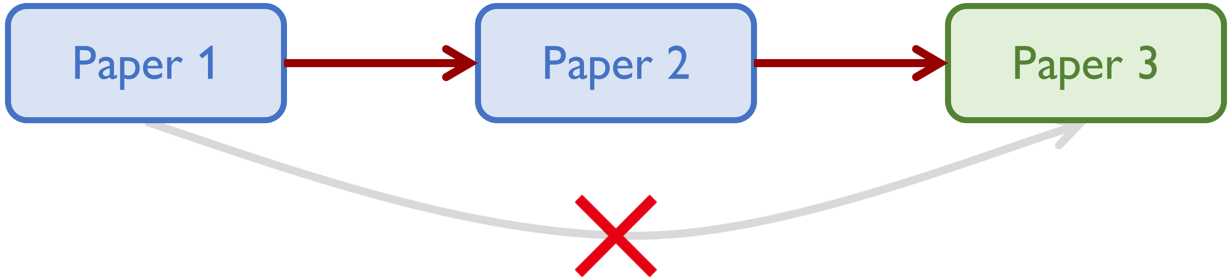
or

$$\frac{\exp(\mathbf{h}_1^T \mathbf{h}_2)}{\exp\left(\mathbf{h}_1^T \mathbf{h}_2\right) + \exp(\mathbf{h}_1^T \mathbf{h}_3)}$$

Hard Negative Samples – SPECTER

- We need to find challenging cases of "Paper 3" so that the model can be improved through contrastive learning.

- The strategy of SPECTER
  - If Paper 1 cites Paper 2, and Paper 2 cites Paper 3, but Paper 1 does not cite Paper 3, then Paper 3 is a hard negative.

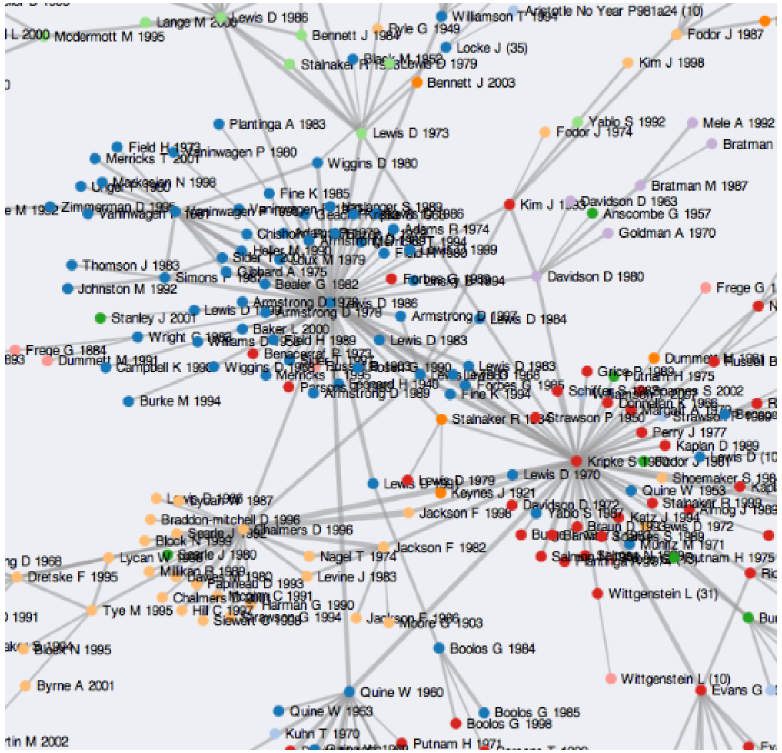


  - Combination of easy and hard negatives: 60% easy + 40% hard

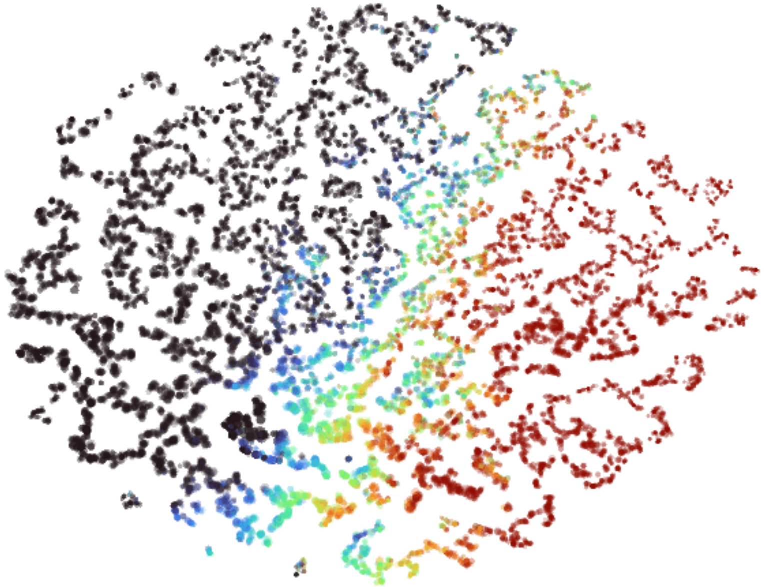
SPECTER: Document-Level Representation Learning using Citation-Informed Transformers. ACL 2020.

Hard Negative Samples – SciNCL

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



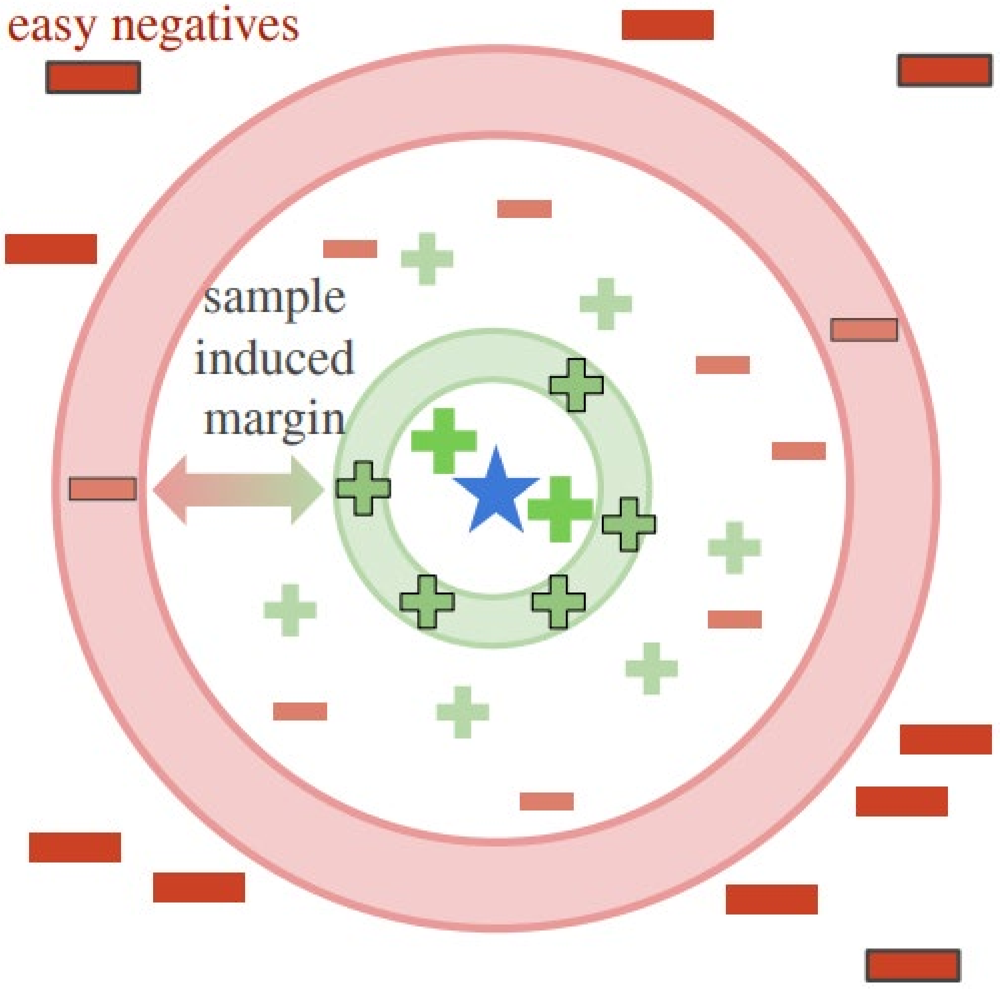
Mapping each node to a vector using graph information only



Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. EMNLP 2022.

Hard Negative Samples – SciNCL

- ★ : query (Paper 1)
- ✚ : easy positive (should NOT be used as Paper 2)
- ✙ : hard positive (should be used as Paper 2)
- confusing area (should NOT be used as Paper 2 or Paper 3)
- ▭ : hard negative (should be used as Paper 3)
- 🟥 : easy negative

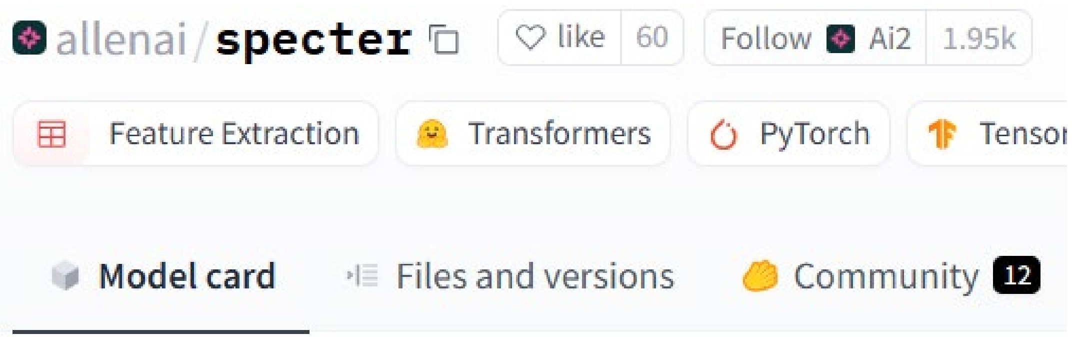


Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. EMNLP 2022.

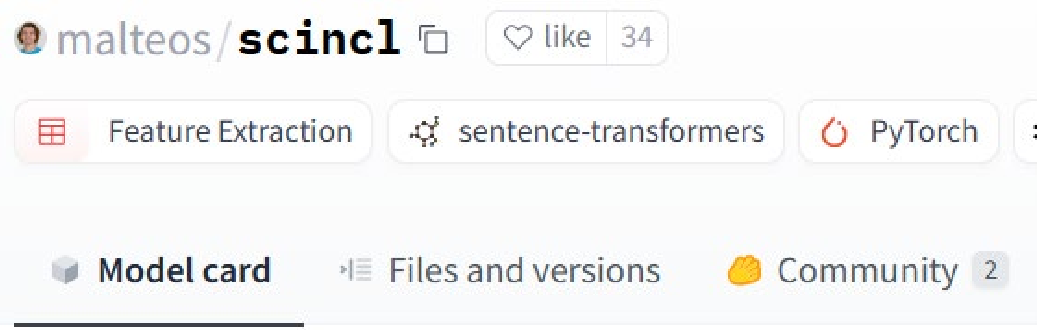
More Details of SPECTER and SciNCL

- Architecture: the same as BERT-base (12-layer Transformer encoders, 110M parameters)
- Pre-training Data: 676K triplets of (query, positive, negative)
- Continue pre-training SciBERT using contrastive learning only

https://huggingface.co/allenai/specter



https://huggingface.co/malteos/scincl

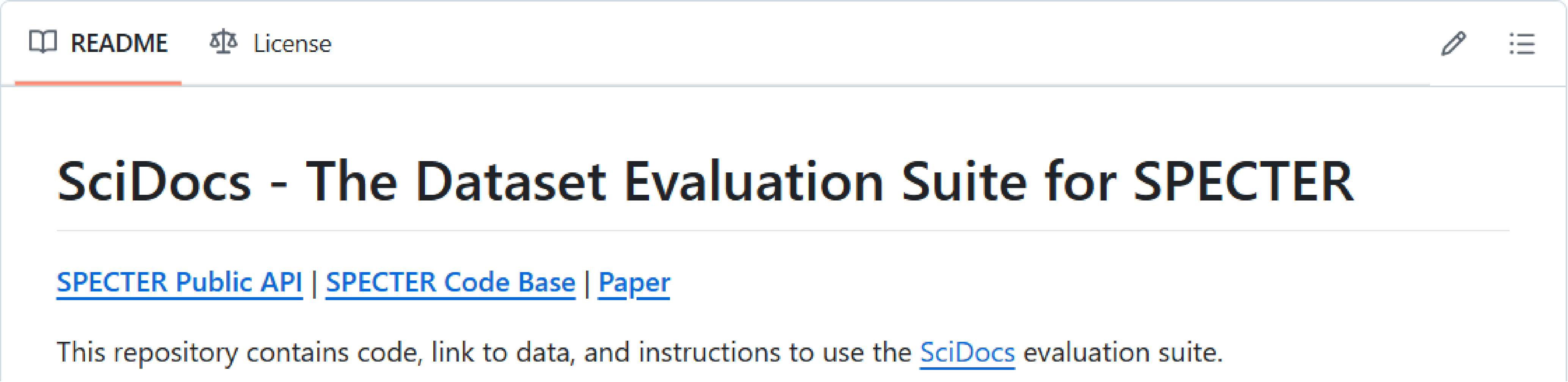


SPECTER: Document-Level Representation Learning using Citation-Informed Transformers. ACL 2020.
Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. EMNLP 2022.

Dataset for Evaluating SPECTER and SciNCL

- The SciDocs benchmark

https://github.com/allenai/scidocs



SciDocs - The Dataset Evaluation Suite for SPECTER

SPECTER Public API | SPECTER Code Base | Paper

This repository contains code, link to data, and instructions to use the SciDocs evaluation suite.

- Citation Prediction: Given a query paper and 30 candidate papers (5 cited by the query and 25 not cited by the query), rank all cited papers higher than all uncited ones.

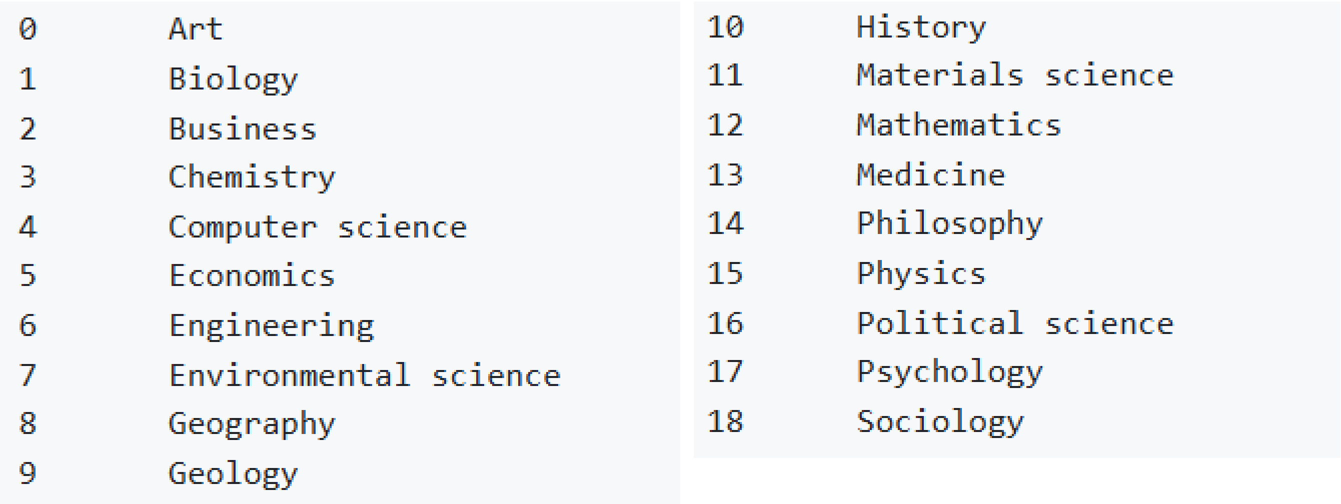
Q2: Can citation information help an LLM with other tasks?

- The SciDocs benchmark
  - Citation
  - Co-Citation: Predict if two papers are frequently cited together.
  - Co-View: Predict if two papers' abstract pages (on Semantic Scholar) are frequently viewed in a single browsing session by users.
  - Co-Read: Predict if two papers' PDF pages (on Semantic Scholar) are frequently viewed in a single browsing session by users.
  - Recommendation: On each paper's abstract page, Semantic Scholar will show some similar papers. Predict which papers are more likely to be clicked by the user.

Q2: Can citation information help an LLM with other tasks?

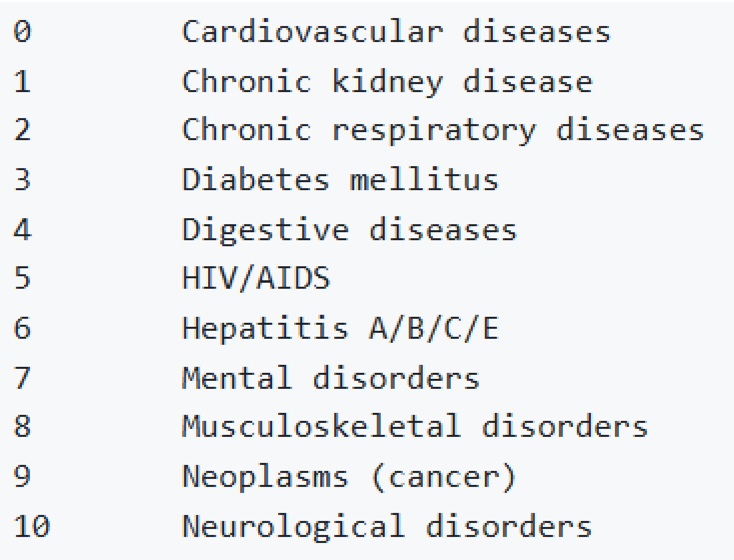
- The SciDocs benchmark
  - “Proximity” Prediction: Citation, Co-Citation, Co-View, Co-Read, Recommendation
  - Classification: MAG (19 classes), MeSH (11 classes)
    - Train an SVM using labeled training data

MAG label space



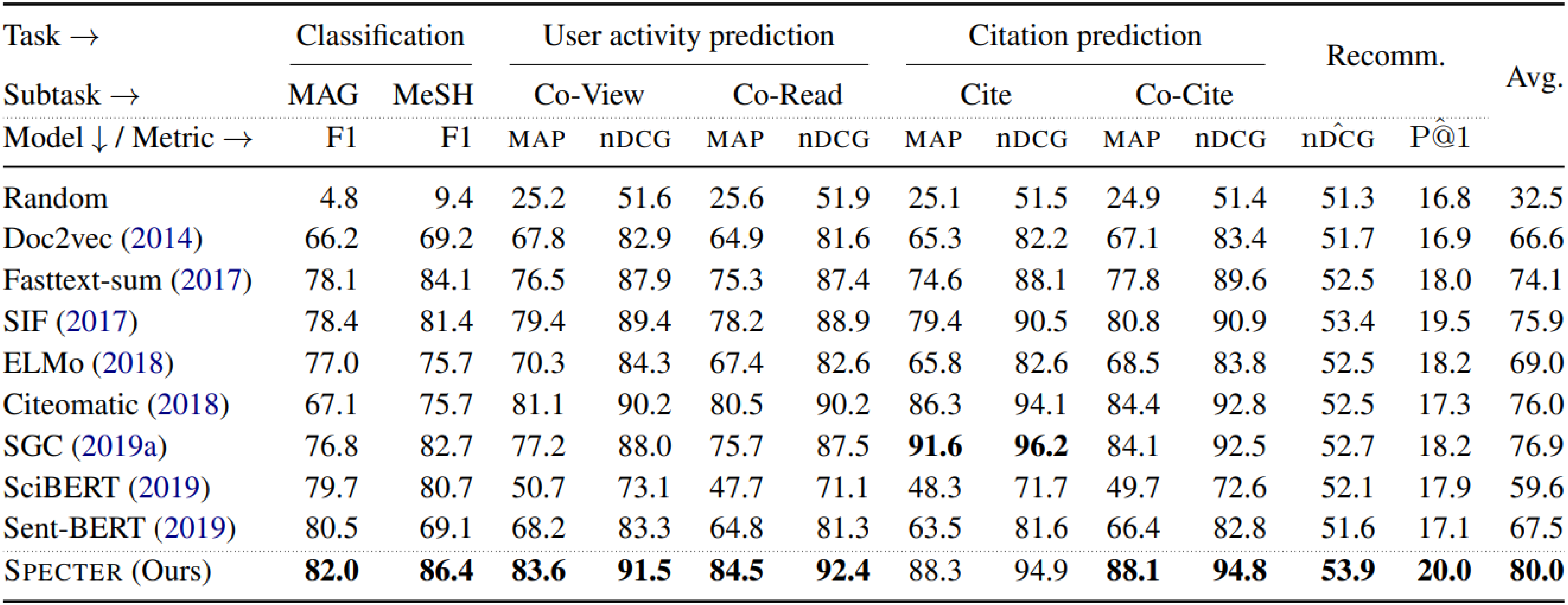
0	Art
1	Biology
2	Business
3	Chemistry
4	Computer science
5	Economics
6	Engineering
7	Environmental science
8	Geography
9	Geology
10	History
11	Materials science
12	Mathematics
13	Medicine
14	Philosophy
15	Physics
16	Political science
17	Psychology
18	Sociology

MeSH label space



0	Cardiovascular diseases
1	Chronic kidney disease
2	Chronic respiratory diseases
3	Diabetes mellitus
4	Digestive diseases
5	HIV/AIDS
6	Hepatitis A/B/C/E
7	Mental disorders
8	Musculoskeletal disorders
9	Neoplasms (cancer)
10	Neurological disorders

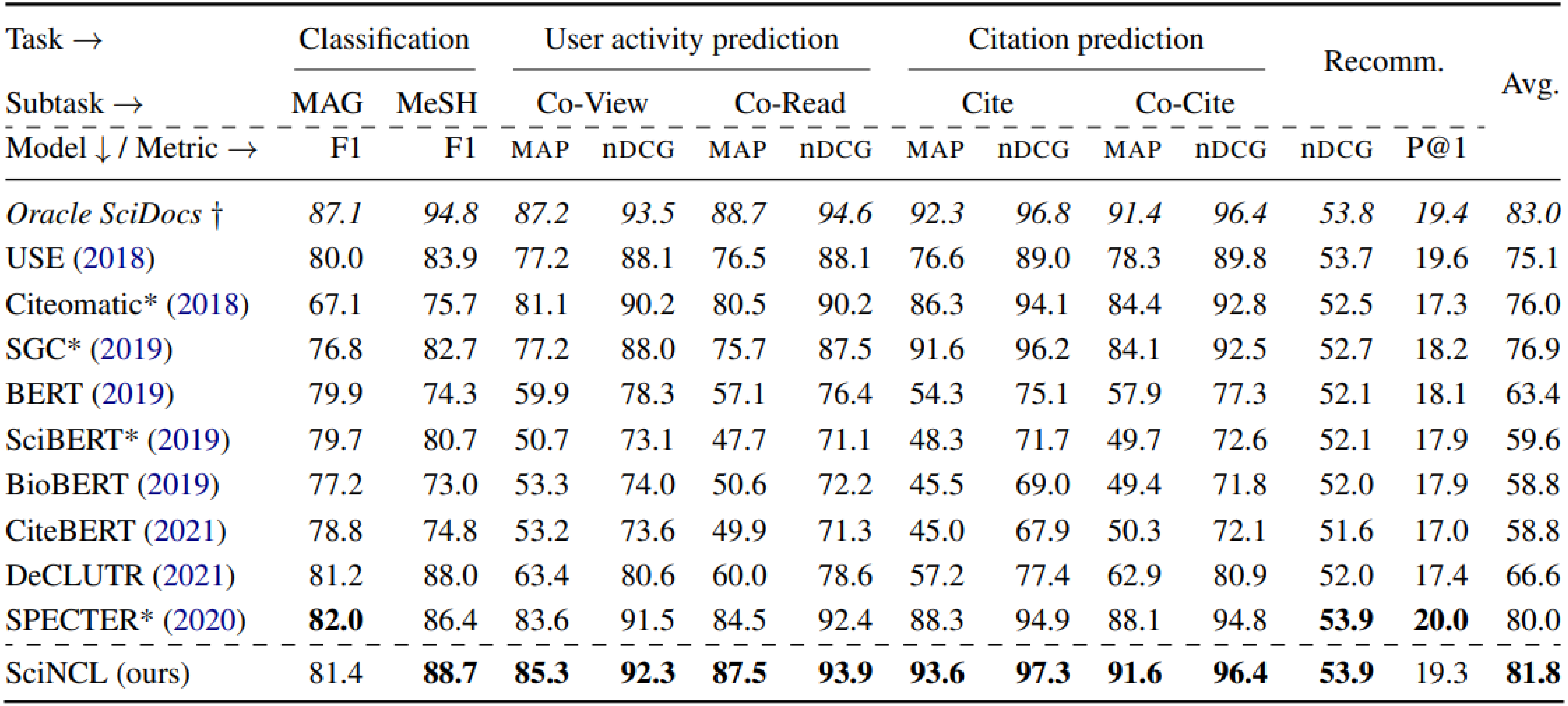
Performance of SPECTER



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	nDĈG	P@̂1	
Random	4.8	9.4	25.2	51.6	25.6	51.9	25.1	51.5	24.9	51.4	51.3	16.8	32.5
Doc2vec (2014)	66.2	69.2	67.8	82.9	64.9	81.6	65.3	82.2	67.1	83.4	51.7	16.9	66.6
Fasttext-sum (2017)	78.1	84.1	76.5	87.9	75.3	87.4	74.6	88.1	77.8	89.6	52.5	18.0	74.1
SIF (2017)	78.4	81.4	79.4	89.4	78.2	88.9	79.4	90.5	80.8	90.9	53.4	19.5	75.9
ELMo (2018)	77.0	75.7	70.3	84.3	67.4	82.6	65.8	82.6	68.5	83.8	52.5	18.2	69.0
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 (2019a)	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
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
Sent-BERT (2019)	80.5	69.1	68.2	83.3	64.8	81.3	63.5	81.6	66.4	82.8	51.6	17.1	67.5
SPECTER (Ours)	**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**

SPECTER: Document-Level Representation Learning using Citation-Informed Transformers. ACL 2020.

Performance of SciNCL

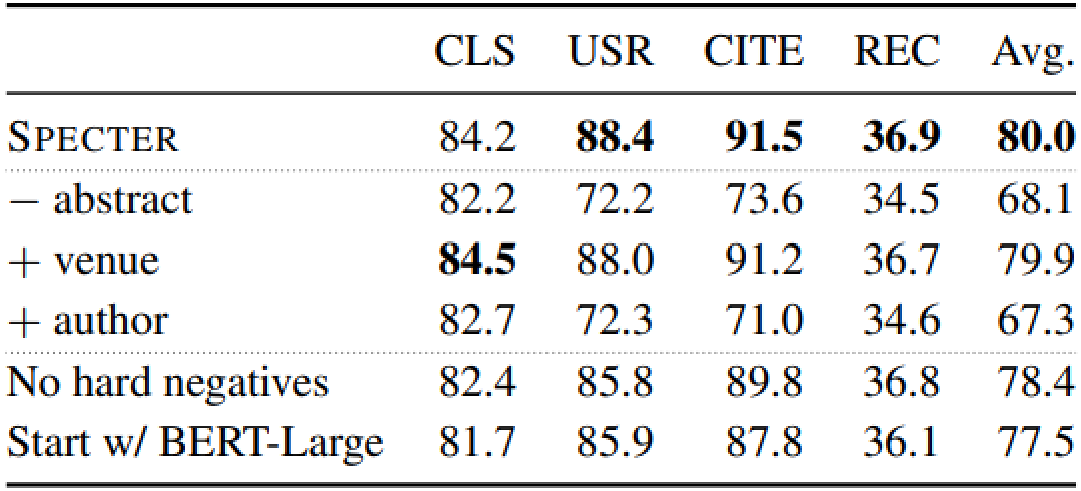


Task →	Classification		User activity prediction				Citation prediction				Recomm.		Avg.
Subtask →	MAG	MeSH	Co-View		Co-Read		Cite		Co-Cite				
Model ↓ / Metric →	F1	F1	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	nDCG	P@1	
Oracle SciDocs †	*87.1*	*94.8*	*87.2*	*93.5*	*88.7*	*94.6*	*92.3*	*96.8*	*91.4*	*96.4*	*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**

Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. EMNLP 2022.

More Experiments in the SPECTER Paper

- What if we do not use any hard negative examples?
  - Harmful!
- What if we feed venue or author information together with paper text into the encoder?
  - Author names are consistently harmful (because the model is never trained to encode person names); venue names only help classification.



	CLS	USR	CITE	REC	Avg.
SPECTER	84.2	**88.4**	**91.5**	**36.9**	**80.0**
− abstract	82.2	72.2	73.6	34.5	68.1
+ venue	**84.5**	88.0	91.2	36.7	79.9
+ author	82.7	72.3	71.0	34.6	67.3
No hard negatives	82.4	85.8	89.8	36.8	78.4
Start w/ BERT-Large	81.7	85.9	87.8	36.1	77.5

SPECTER: Document-Level Representation Learning using Citation-Informed Transformers. ACL 2020.

Take-Away Messages

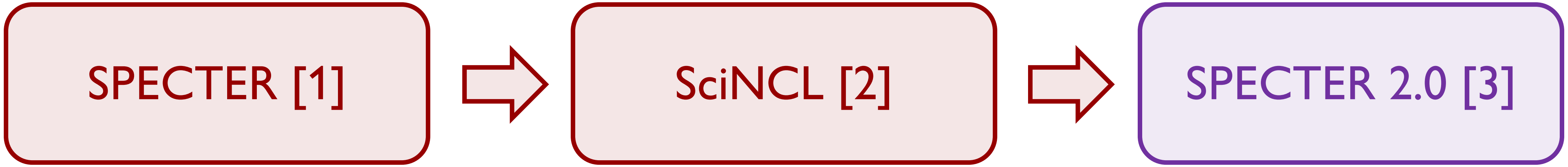
- Citation prediction complements masked language modeling in scientific LLM pre-training. It helps downstream tasks including not only citation prediction but also classification and other types of "proximity" prediction.

- Hard negatives/positives are important in contrastive learning.

- Unsolved issues
  - How to better utilize venue and author information? *OAG-BERT: Towards a Unified Backbone Language Model for Academic Knowledge Services*. KDD 2022.
  - All examined tasks focus on the representation of the entire paper. Can SPECTER and SciNCL outperform SciBERT in named entity recognition? Why (not)?

Two Questions Related to Citations

- Question 1 (enhanced version): How to train an LLM to perform multiple tasks (e.g., citation prediction and classification) simultaneously?
- Question 2 (enhanced version): Can these tasks help an LLM with unseen tasks?



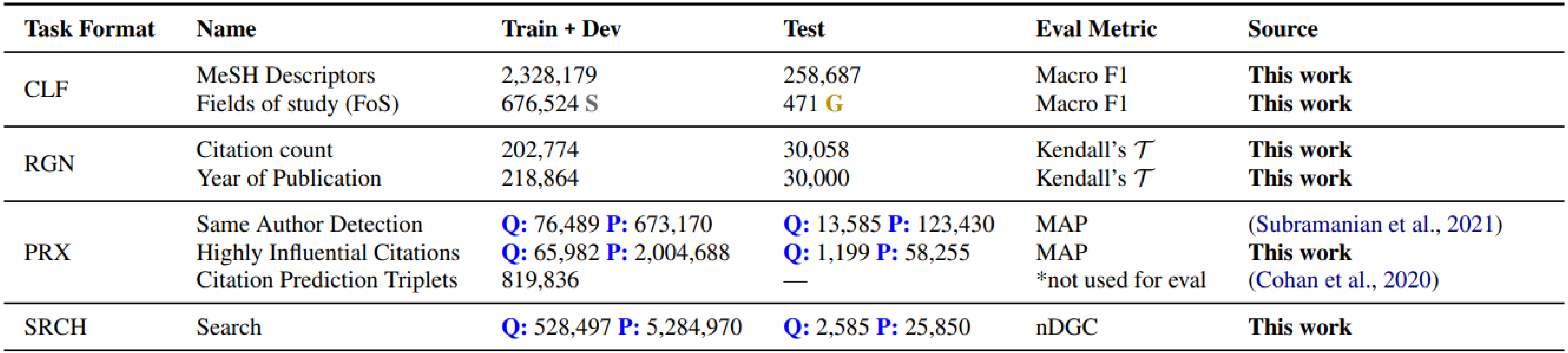
SPECTER [1] → SciNCL [2] → SPECTER 2.0 [3]

[1] *SPECTER: Document-Level Representation Learning using Citation-Informed Transformers*. ACL 2020.

[2] *Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings*. EMNLP 2022.

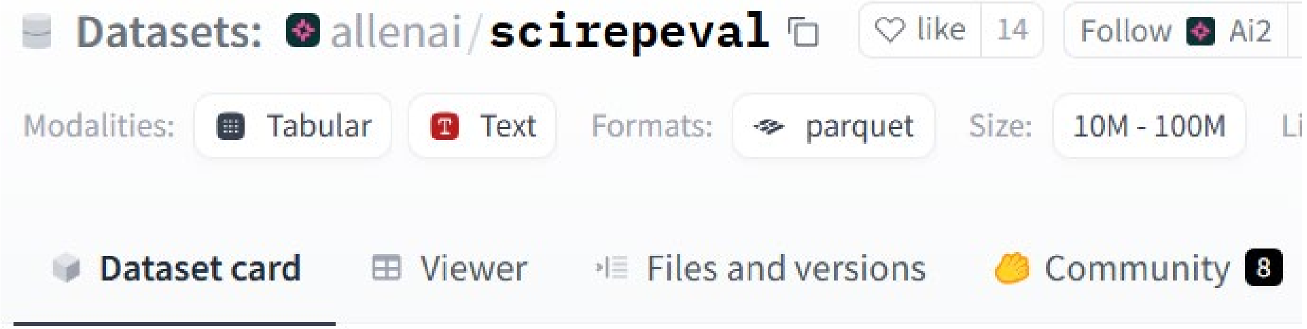
[3] *SciRepEval: A Multi-Format Benchmark for Scientific Document Representations*. EMNLP 2023.

Pre-training Data of SPECTER 2.0 – SciRepEval



Task Format	Name	Train + Dev	Test	Eval Metric	Source
CLF	MeSH Descriptors	2,328,179	258,687	Macro F1	**This work**
	Fields of study (FoS)	676,524 S	471 G	Macro F1	**This work**
RGN	Citation count	202,774	30,058	Kendall's $\mathcal{T}$	**This work**
	Year of Publication	218,864	30,000	Kendall's $\mathcal{T}$	**This work**
PRX	Same Author Detection	**Q:** 76,489 **P:** 673,170	**Q:** 13,585 **P:** 123,430	MAP	(Subramanian et al., 2021)
	Highly Influential Citations	**Q:** 65,982 **P:** 2,004,688	**Q:** 1,199 **P:** 58,255	MAP	**This work**
	Citation Prediction Triplets	819,836	—	*not used for eval	(Cohan et al., 2020)
SRCH	Search	**Q:** 528,497 **P:** 5,284,970	**Q:** 2,585 **P:** 25,850	nDGC	**This work**

https://huggingface.co/datasets/allenai/scirepeval



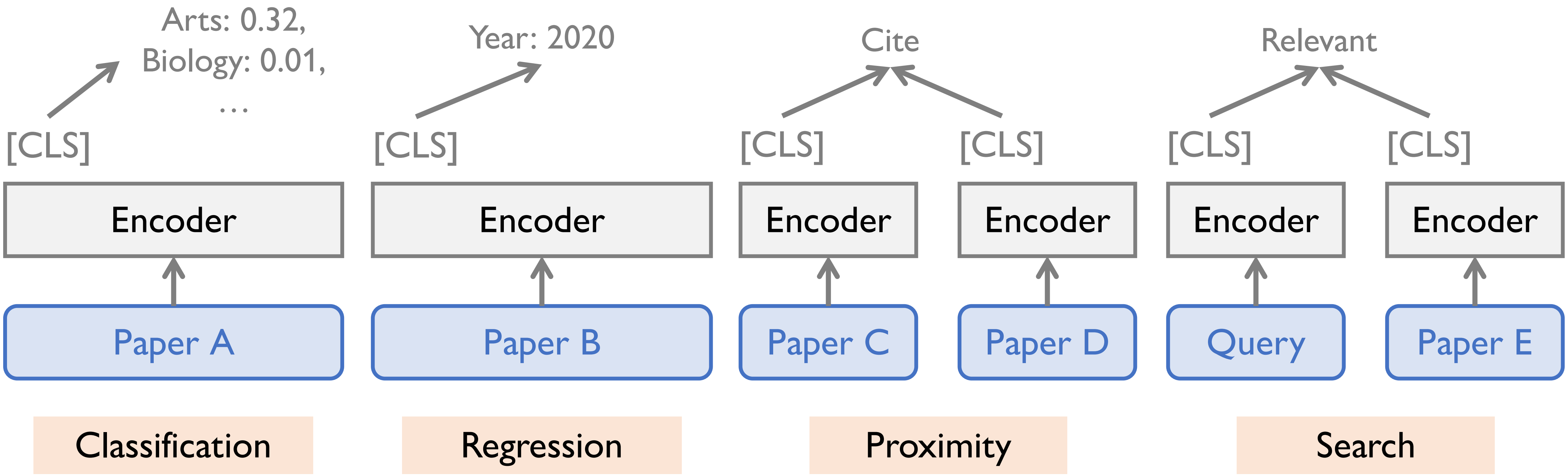
SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. EMNLP 2023.

Pre-training Data of SPECTER 2.0 – SciRepEval

- 4 types of tasks
  - Classification (CLF): predict the MeSH or MAG labels of a paper
  - Regression (RGN): predict the citation count or the publication year of a paper
  - “Proximity” Prediction (PRX)
    - Citation Prediction: predict if one paper cites the other
    - Highly Influential Citation Prediction: predict if one paper frequently cites the other in its text
    - Same Author Detection: predict if two papers are written by the same author
  - Search (SRCH): given a query and a list of papers, predict which papers are relevant to the query (derived from Semantic Scholar search logs)

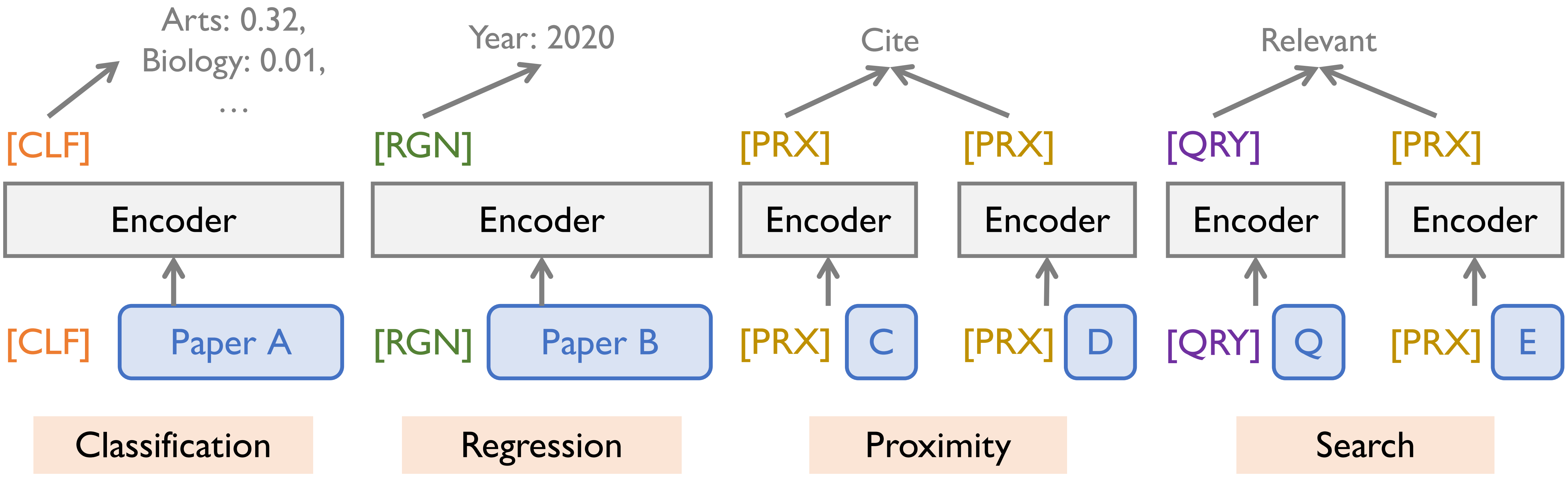
How to pre-train an LLM with multiple tasks?

- Vanilla version



How to pre-train an LLM with multiple tasks?

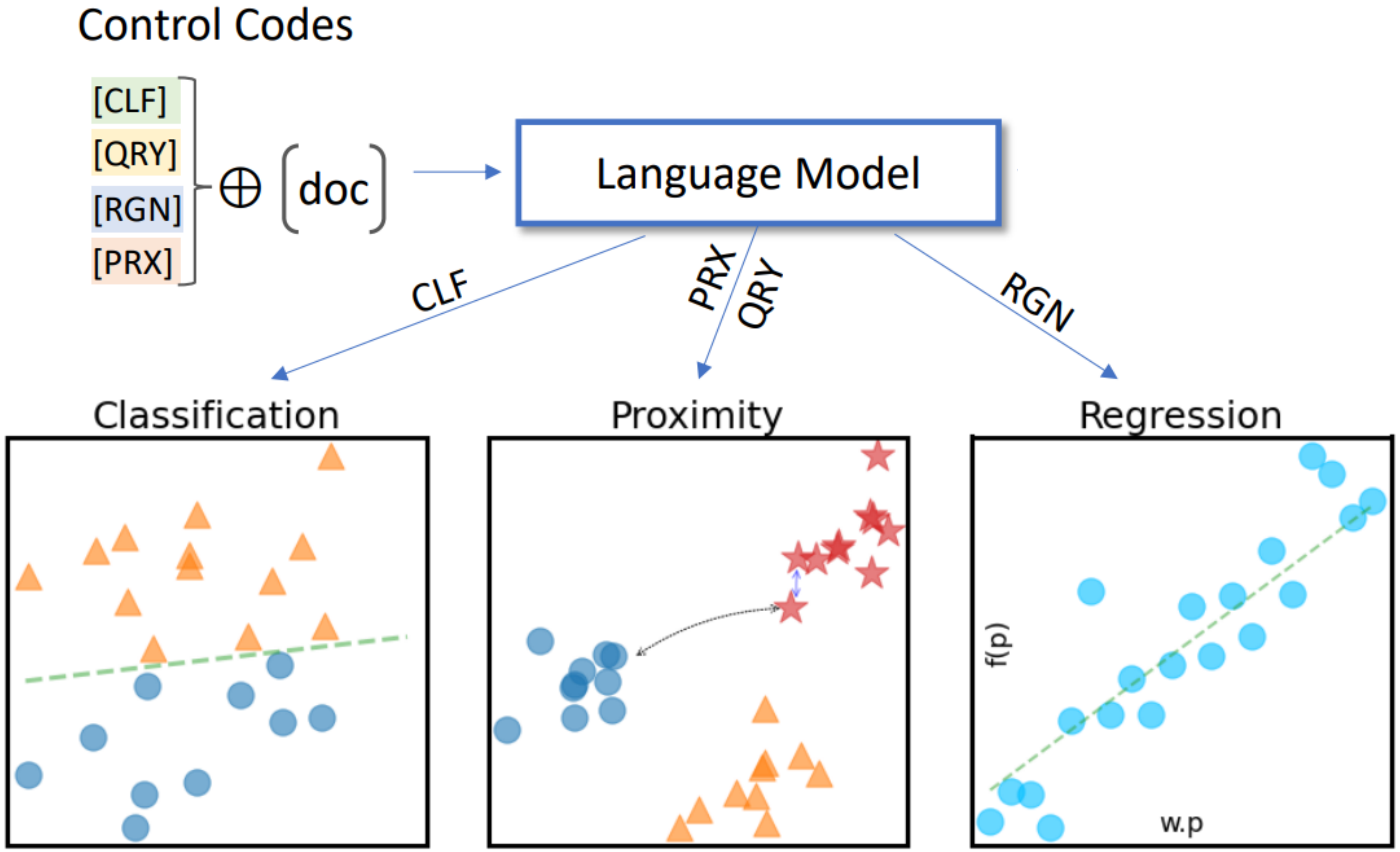
- Trick 1: Control Codes



How to pre-train an LLM with multiple tasks?

- Trick 1: Control Codes

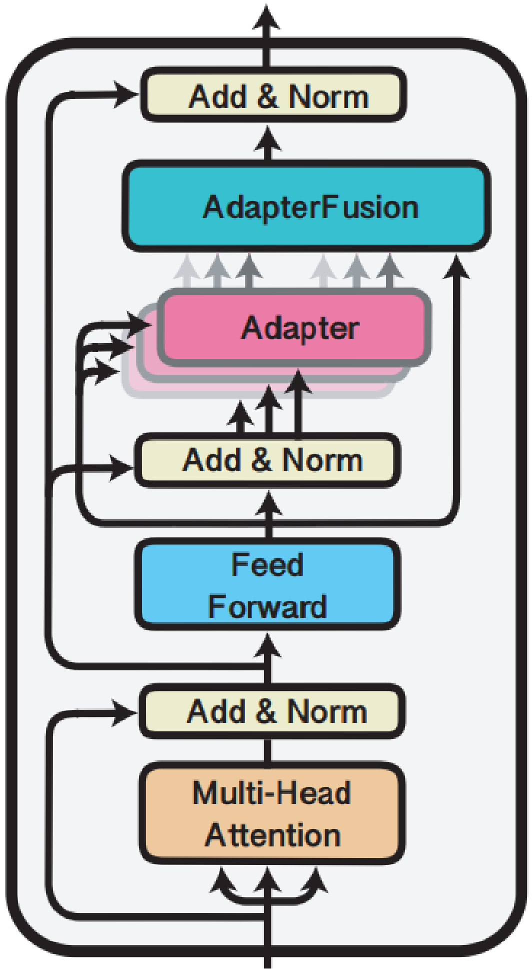
Motivation: You need different embedding spaces when performing different downstream tasks.



How to pre-train an LLM with multiple tasks?

- Trick 1: Control Codes - All tasks share the same architecture. We get different embeddings of a paper by slightly changing the input.

- Trick 2: Adapters - Different tasks have their shared parameters and task-specific parameters.
  - Shared parameters: Multi-Head Attention and Feed Forward; representing task commonality
  - Task-specific parameters: Adapter; representing task specificity
    - If the model is performing classification, the data will go through the "classification" adapter.

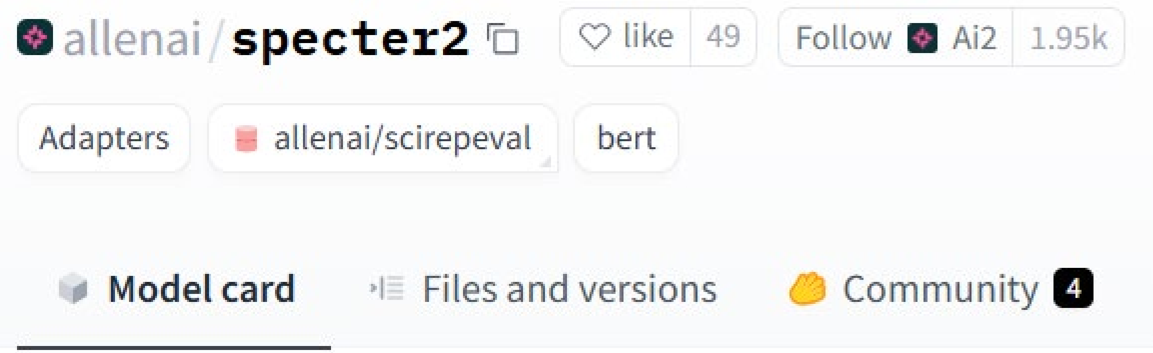


Parameter-Efficient Transfer Learning for NLP. ICML 2019.
AdapterFusion: Non-Destructive Task Composition for Transfer Learning. EACL 2021.

More Details of SPECTER 2.0

- Architecture: 12-layer × (Transformer + Adapters), 113M parameters
- Continue pre-training SciBERT using classification, regression, proximity prediction, and search

https://huggingface.co/allenai/specter2

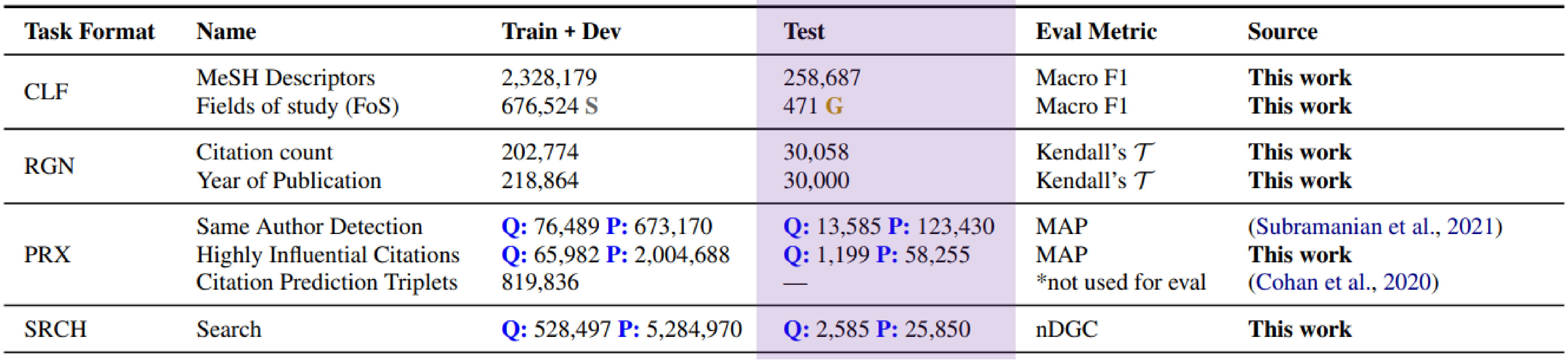


allenai / specter2

SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. EMNLP 2023.

Tasks for Evaluating SPECTER 2.0

- “In-train” tasks

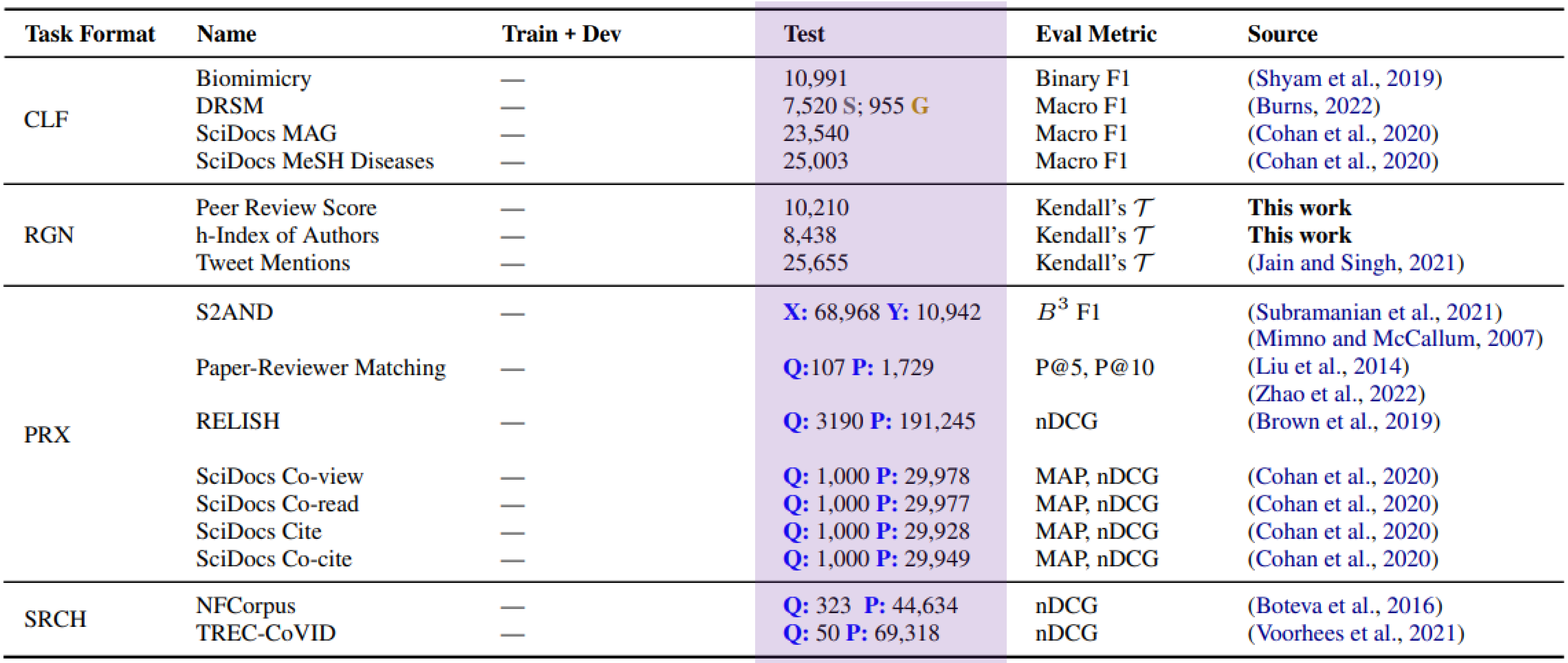


Task Format	Name	Train + Dev	Test	Eval Metric	Source
CLF	MeSH Descriptors	2,328,179	258,687	Macro F1	**This work**
	Fields of study (FoS)	676,524 **S**	471 **G**	Macro F1	**This work**
RGN	Citation count	202,774	30,058	Kendall’s $\mathcal{T}$	**This work**
	Year of Publication	218,864	30,000	Kendall’s $\mathcal{T}$	**This work**
PRX	Same Author Detection	**Q:** 76,489 **P:** 673,170	**Q:** 13,585 **P:** 123,430	MAP	(Subramanian et al., 2021)
	Highly Influential Citations	**Q:** 65,982 **P:** 2,004,688	**Q:** 1,199 **P:** 58,255	MAP	**This work**
	Citation Prediction Triplets	819,836	—	*not used for eval	(Cohan et al., 2020)
SRCH	Search	**Q:** 528,497 **P:** 5,284,970	**Q:** 2,585 **P:** 25,850	nDGC	**This work**

SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. EMNLP 2023.

Tasks for Evaluating SPECTER 2.0

- “Out-of-train” tasks



Task Format	Name	Train + Dev	Test	Eval Metric	Source
CLF	Biomimicry	—	10,991	Binary F1	(Shyam et al., 2019)
	DRSM	—	7,520 S; 955 G	Macro F1	(Burns, 2022)
	SciDocs MAG	—	23,540	Macro F1	(Cohan et al., 2020)
	SciDocs MeSH Diseases	—	25,003	Macro F1	(Cohan et al., 2020)
RGN	Peer Review Score	—	10,210	Kendall’s τ	**This work**
	h-Index of Authors	—	8,438	Kendall’s τ	**This work**
	Tweet Mentions	—	25,655	Kendall’s τ	(Jain and Singh, 2021)
PRX	S2AND	—	**X:** 68,968 **Y:** 10,942	B^3 F1	(Subramanian et al., 2021)
	Paper-Reviewer Matching	—	**Q:**107 **P:** 1,729	P@5, P@10	(Mimno and McCallum, 2007) (Liu et al., 2014) (Zhao et al., 2022)
	RELISH	—	**Q:** 3190 **P:** 191,245	nDCG	(Brown et al., 2019)
	SciDocs Co-view	—	**Q:** 1,000 **P:** 29,978	MAP, nDCG	(Cohan et al., 2020)
	SciDocs Co-read	—	**Q:** 1,000 **P:** 29,977	MAP, nDCG	(Cohan et al., 2020)
	SciDocs Cite	—	**Q:** 1,000 **P:** 29,928	MAP, nDCG	(Cohan et al., 2020)
	SciDocs Co-cite	—	**Q:** 1,000 **P:** 29,949	MAP, nDCG	(Cohan et al., 2020)
SRCH	NFCorpus	—	**Q:** 323 **P:** 44,634	nDCG	(Boteva et al., 2016)
	TREC-CoVID	—	**Q:** 50 **P:** 69,318	nDCG	(Voorhees et al., 2021)

SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. EMNLP 2023.

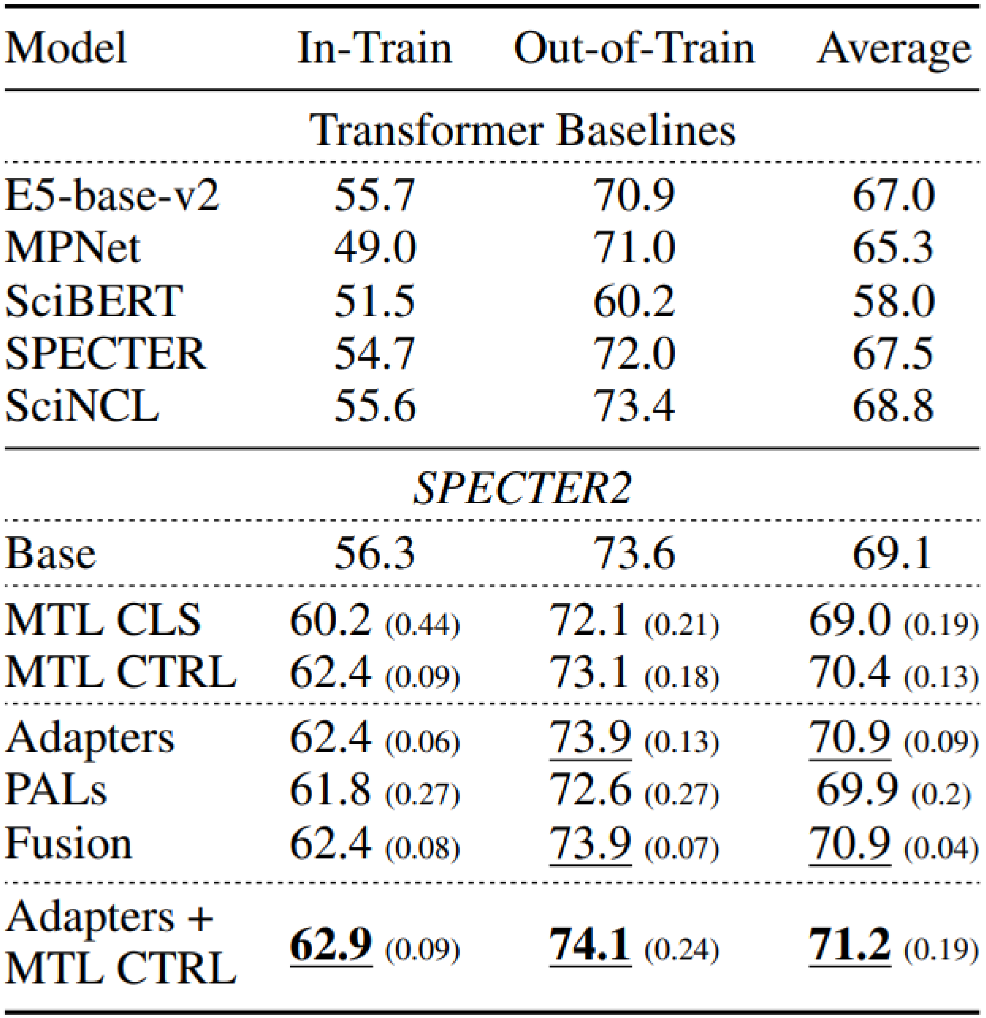
Tasks for Evaluating SPECTER 2.0

- “Out-of-train” tasks
  - Classification
    - Biomimicry: predict if a paper is related to biomimicry
    - DRSM: predict which aspect of rare diseases a paper deals with (6 aspects in total)
  - Regression
    - Peer Review Score: predict the average score each ICLR submission gets (between 1 and 10)
    - h-Index of Authors: given a paper, predict the maximum h-index of any of the authors
    - Tweet Mentions: given a paper, predict how many times it is mentioned and retweeted

Tasks for Evaluating SPECTER 2.0

- “Out-of-train” tasks
  - Proximity Prediction
    - S2AND (Author Name Disambiguation): given many papers written by many authors with the same name, cluster the papers according to their authors
    - Paper-Reviewer Matching: given a submission and a list of candidate reviewers (with their previously published papers), rank the reviewers according to their expertise to review the submission
  - Search
    - NFCorpus: given a query and a list of papers (about nutrition facts), rank the papers according to their relevance to the query
    - TREC-COVID: given a query and a list of papers (about COVID-19), rank the papers according to their relevance to the query

Performance of SPECTER 2.0

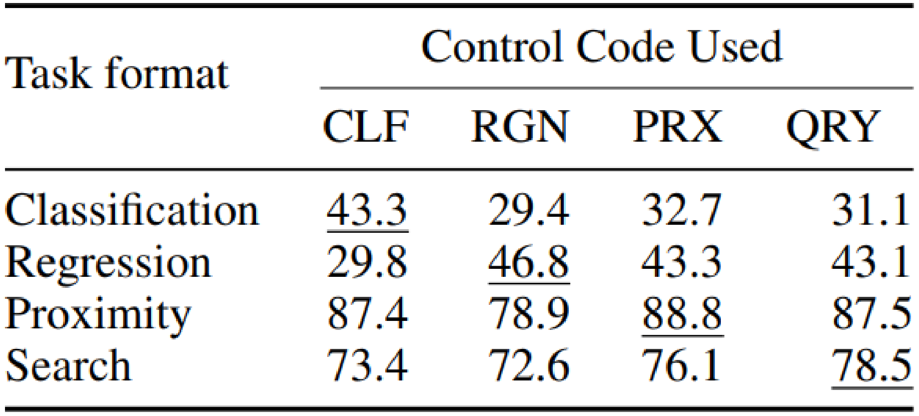


Model	In-Train	Out-of-Train	Average
Transformer Baselines			
E5-base-v2	55.7	70.9	67.0
MPNet	49.0	71.0	65.3
SciBERT	51.5	60.2	58.0
SPECTER	54.7	72.0	67.5
SciNCL	55.6	73.4	68.8
SPECTER2			
Base	56.3	73.6	69.1
MTL CLS	60.2 (0.44)	72.1 (0.21)	69.0 (0.19)
MTL CTRL	62.4 (0.09)	73.1 (0.18)	70.4 (0.13)
Adapters	62.4 (0.06)	73.9 (0.13)	70.9 (0.09)
PALs	61.8 (0.27)	72.6 (0.27)	69.9 (0.2)
Fusion	62.4 (0.08)	73.9 (0.07)	70.9 (0.04)
Adapters + MTL CTRL	**62.9** (0.09)	**74.1** (0.24)	**71.2** (0.19)

Using [CLS] Only
Using Control Codes

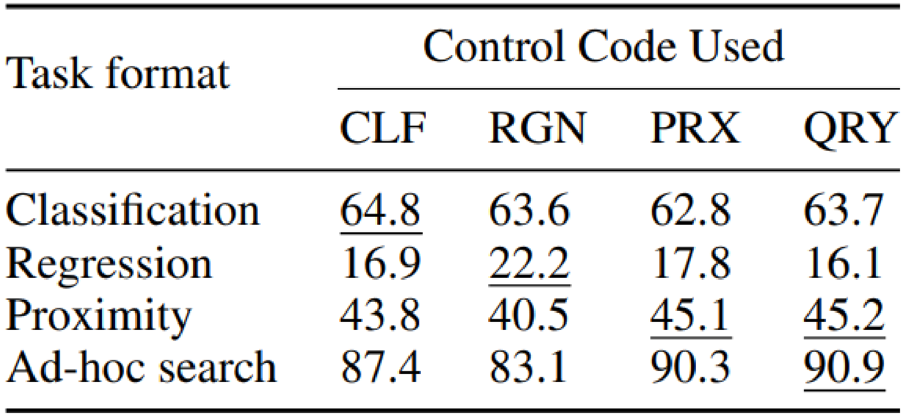
Using Adapters
Variant of Adapters
Variant of Adapters

Using Adapters +
Control Codes



Task format	Control Code Used			
	CLF	RGN	PRX	QRY
Classification	43.3	29.4	32.7	31.1
Regression	29.8	46.8	43.3	43.1
Proximity	87.4	78.9	88.8	87.5
Search	73.4	72.6	76.1	78.5

(a) in-train



Task format	Control Code Used			
	CLF	RGN	PRX	QRY
Classification	64.8	63.6	62.8	63.7
Regression	16.9	22.2	17.8	16.1
Proximity	43.8	40.5	45.1	45.2
Ad-hoc search	87.4	83.1	90.3	90.9

(b) out-of-train

SciRepEval:A Multi-Format Benchmark for Scientific Document Representations. EMNLP 2023.

Take-Away Messages

- Pre-training an LLM using multiple tasks (e.g., classification, regression, citation prediction, search) makes it perform better in both in-train and out-of-train tasks.
    - The motivation is similar to instruction tuning!

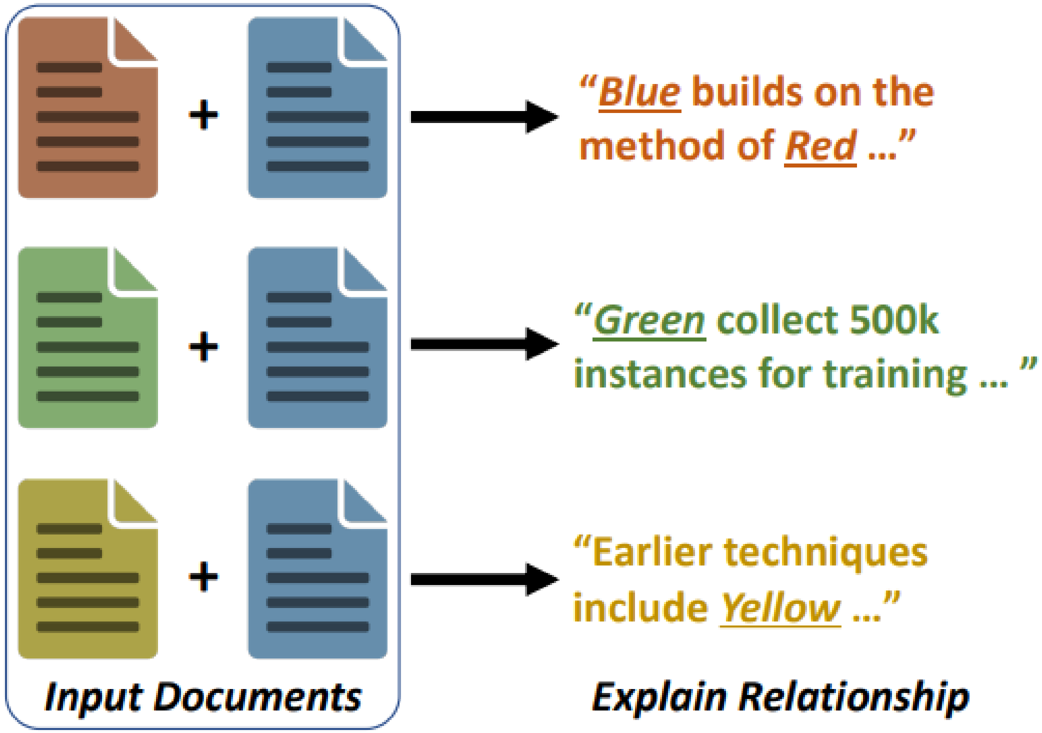
- When performing different tasks, it is better to generate different embeddings for the same text.
    - Control codes: shared architecture + task-specific inputs
    - Adapters: partially shared + partially task-specific architecture

Take-Away Messages

- Drawback:
    - What if we have an entirely new task without training data?
        - We have to choose an existing adapter or an existing code to perform this task.
        - Invent a new control code? Control codes are not natural language instructions. The model can hardly understand it.
        - Use natural language instructions to replace control codes during pre-training?
            - Instruction tuning + encoder-only architectures
            - *Task-aware Retrieval with Instructions*. ACL 2023 Findings.
            - *Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding*. EMNLP 2023 Findings.

What can decoder architectures do with citation information?

- Given two papers (one citing the other), explain the relationship between them.
  - A generative version of citation intent prediction.



"*Blue* builds on the method of *Red* ..."

"*Green* collect 500k instances for training ... "

"Earlier techniques include *Yellow* ..."

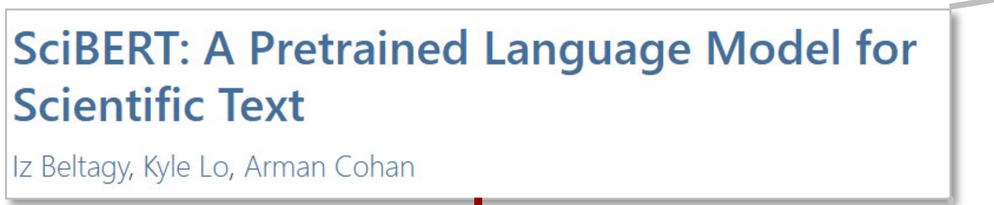
Input Documents

Explain Relationship

Explaining Relationships Between Scientific Documents. ACL 2021.

How to collect data?

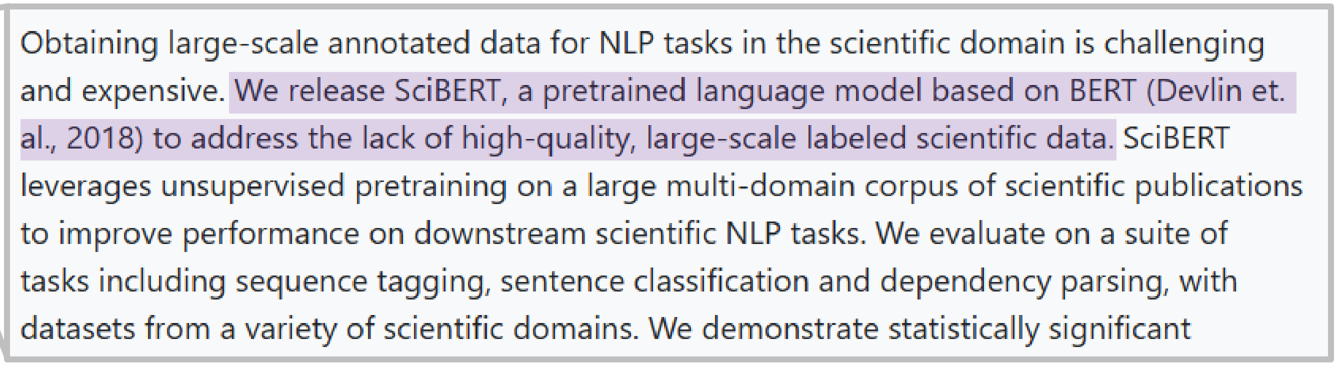
Anchor Sentence



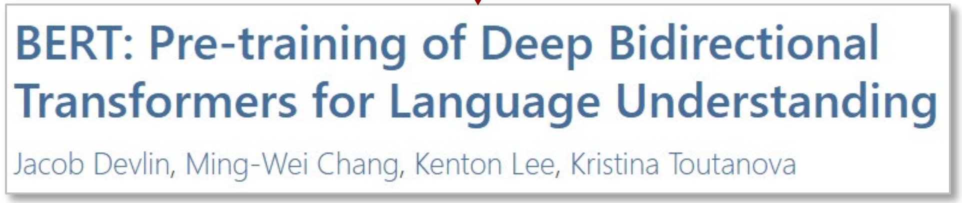
SciBERT: A Pretrained Language Model for Scientific Text

Iz Beltagy, Kyle Lo, Arman Cohan

Principal Paper



Obtaining large-scale annotated data for NLP tasks in the scientific domain is challenging and expensive. We release SciBERT, a pretrained language model based on BERT (Devlin et. al., 2018) to address the lack of high-quality, large-scale labeled scientific data. SciBERT leverages unsupervised pretraining on a large multi-domain corpus of scientific publications to improve performance on downstream scientific NLP tasks. We evaluate on a suite of tasks including sequence tagging, sentence classification and dependency parsing, with datasets from a variety of scientific domains. We demonstrate statistically significant



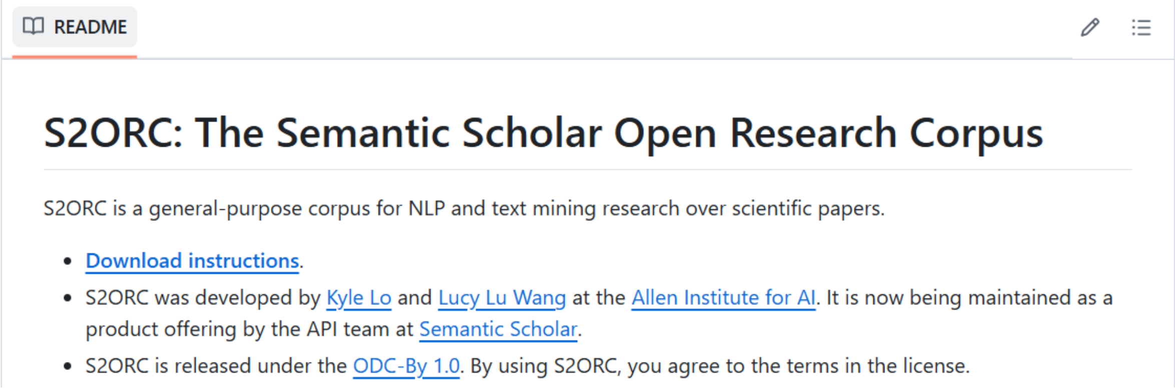
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

Cited Paper

https://github.com/allenai/s2orc

title, abstract, full text, citations, anchor sentences, …

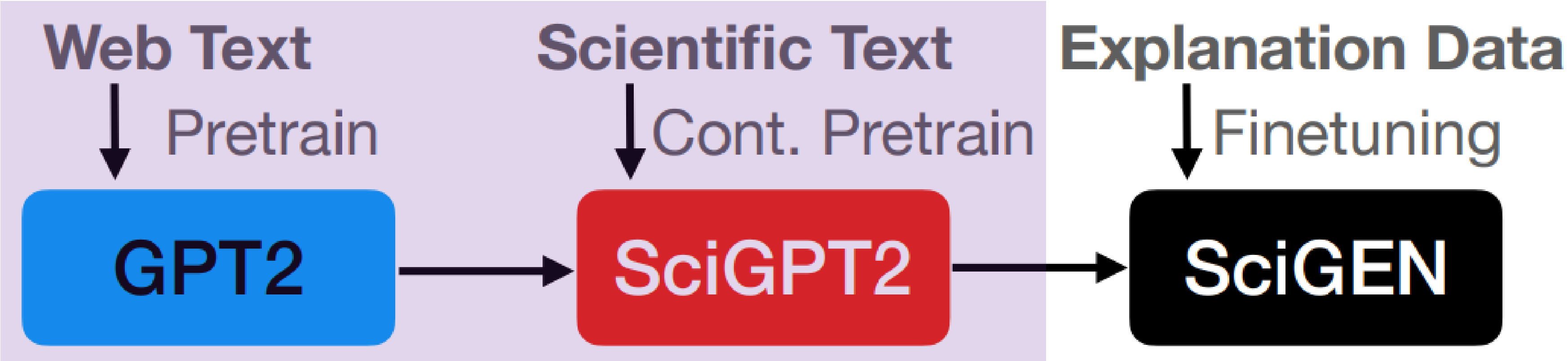


S2ORC: The Semantic Scholar Open Research Corpus

S2ORC is a general-purpose corpus for NLP and text mining research over scientific papers.

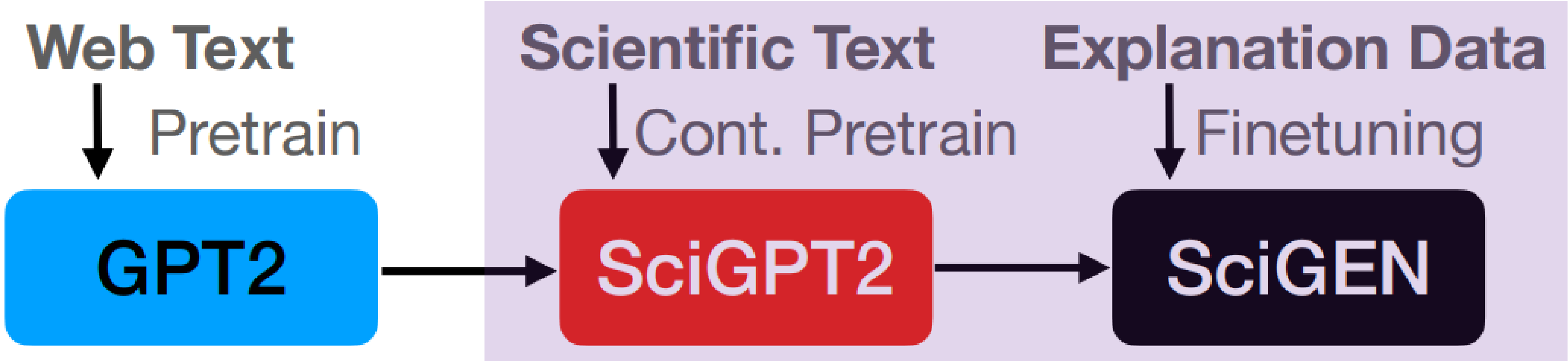
- Download instructions.
- S2ORC was developed by Kyle Lo and Lucy Lu Wang at the Allen Institute for AI. It is now being maintained as a product offering by the API team at Semantic Scholar.
- S2ORC is released under the ODC-By 1.0. By using S2ORC, you agree to the terms in the license.

Roadmap to the Model

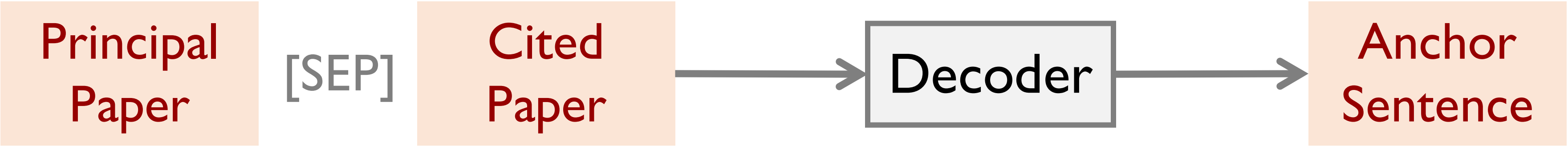


- Step 1: Continue pre-training GPT-2 using unsupervised next token prediction on a large scientific paper corpus

Roadmap to the Model



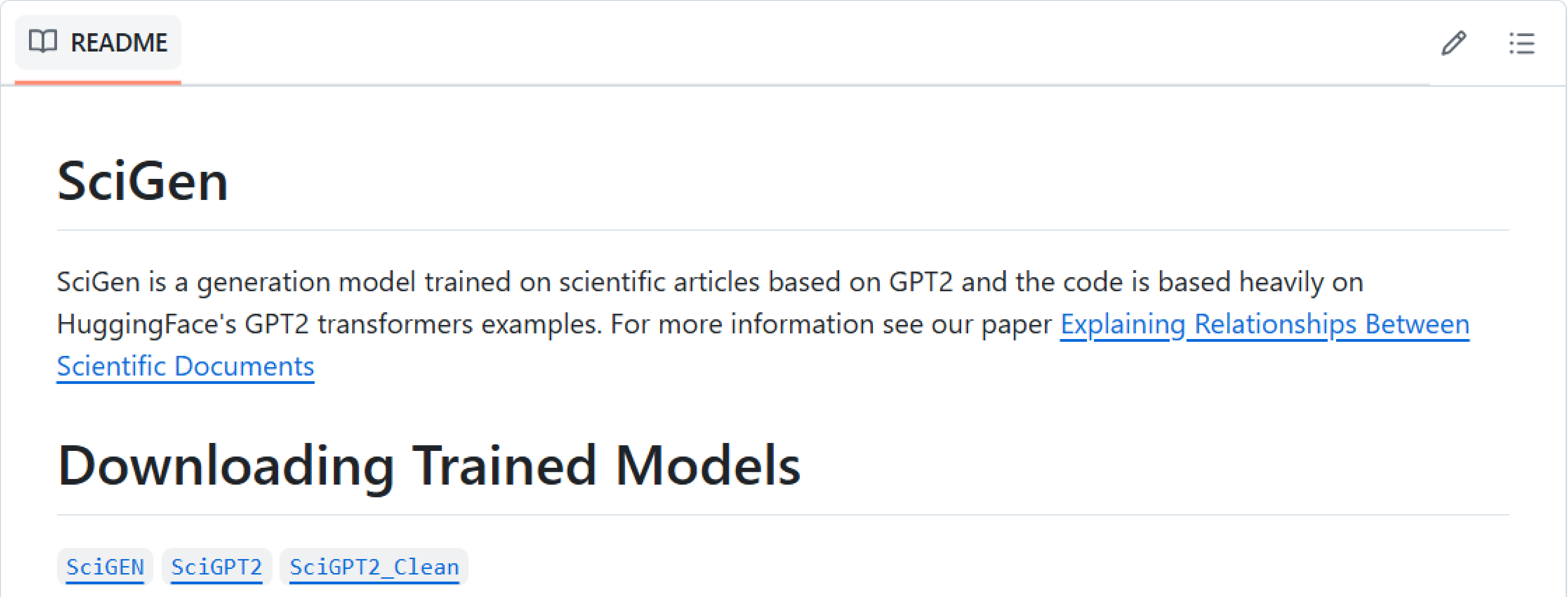
- Step 1: Continue pre-training GPT-2 using unsupervised next token prediction on a large scientific paper corpus
- Step 2: Supervised fine-tuning



More Details of SciGEN

- Architecture: the same as GPT-2-base (12-layer Transformer decoders, 117M parameters)
- Fine-tuning Data: 622K triplets of (principal paper, cited paper, anchor sentence)

https://github.com/Kel-Lu/SciGen



SciGen

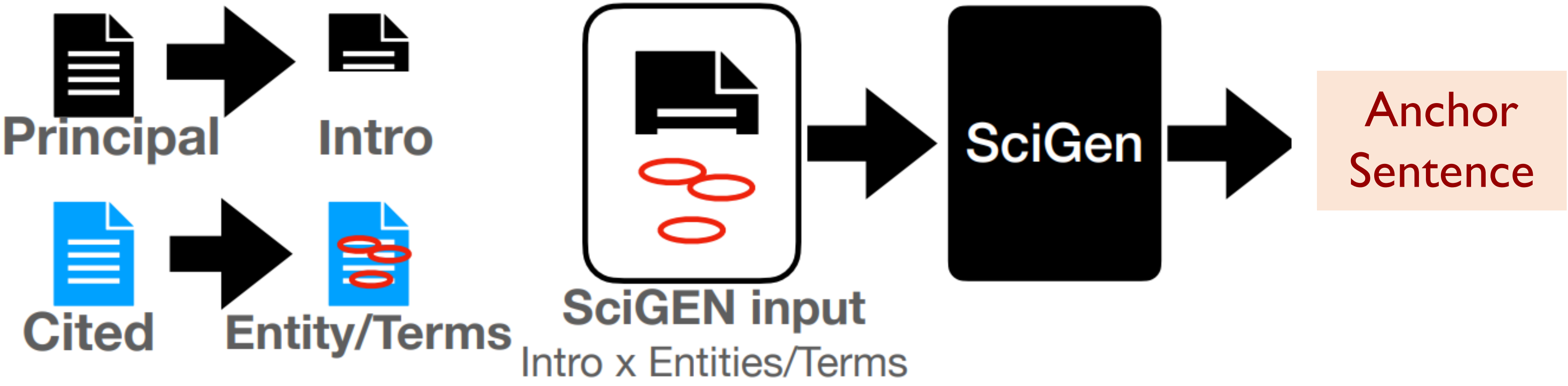
SciGen is a generation model trained on scientific articles based on GPT2 and the code is based heavily on HuggingFace's GPT2 transformers examples. For more information see our paper Explaining Relationships Between Scientific Documents

Downloading Trained Models

SciGEN SciGPT2 SciGPT2_Clean

Performance of SciGEN

Context	BLEU	ACL-BLEU	Rouge-L
principal abs × cited abs	9.82	10.40	8.4
principal intro × cited abs	9.92	11.22	8.7
principal intro × cited intro	9.80	10.54	8.8
principal intro × cited sampled	9.81	10.31	8.7
principal intro × cited tfidf	13.17	16.75	12.0
principal intro × cited entities	13.41	13.42	11.8



Explaining Relationships Between Scientific Documents. ACL 2021.

Take-Away Messages

- Citations are associated with text information (i.e., anchor sentences), making them beyond edges in a graph.
- Such text information can help explain document relationships.

- Keywords extracted by TF-IDF scores are more useful than the abstract/introduction when representing the cited paper as input to the model.
  - Is this observation still true for GPT-3 or even stronger LLMs?

- Drawback
  - Evaluation metrics include BLEU and ROUGE only, which are based on word overlaps between the generated text and the ground-truth text.
  - *BERTScore: Evaluating Text Generation with BERT.* ICLR 2020.
  - *GPTScore: Evaluate as You Desire.* NAACL 2024.

Thank You!

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