Introduction to SciRepEval

And SciRank, a scientific document representation model we are developing using SciRepEval data

BACKGROUND of SciRank

- When scholars search for papers, they care about both the relevance and the authority of the results. Ranking documents by relevance, dense retrieval models have shown good performance in academic search but don’t take into account the authority of scientific papers
- To rank papers by authority, major academic search engines use citation counts as the most important ranking factor for authority. Is there a better measure?

APPROACH

- Our model consists of two parts: language model and a centrality measure.
- For search tasks, we also include Elastic Search which uses BM25, a classical ranking model.

Language Models

- SPECTER
- SPECTER2
- SciNCL
- LinkBERT-Large

SPECTER and SPECTER 2

- SPECTER is a model for learning task-independent representations of academic papers. It is based on SciBERT and incorporates inter-document information through citations. SPECTER pretrains the model on a large corpus of citations and uses a triplet margin loss function for training, which encourages the model to output representations that are more similar for papers that share a citation link than for those that do not.
- SPECTER 2 is the successor to SPECTER. It was trained in the same way as SPECTER but with more data. It performs slightly better than SPECTER based on Allen AI's experiments and my replication of their experiments.

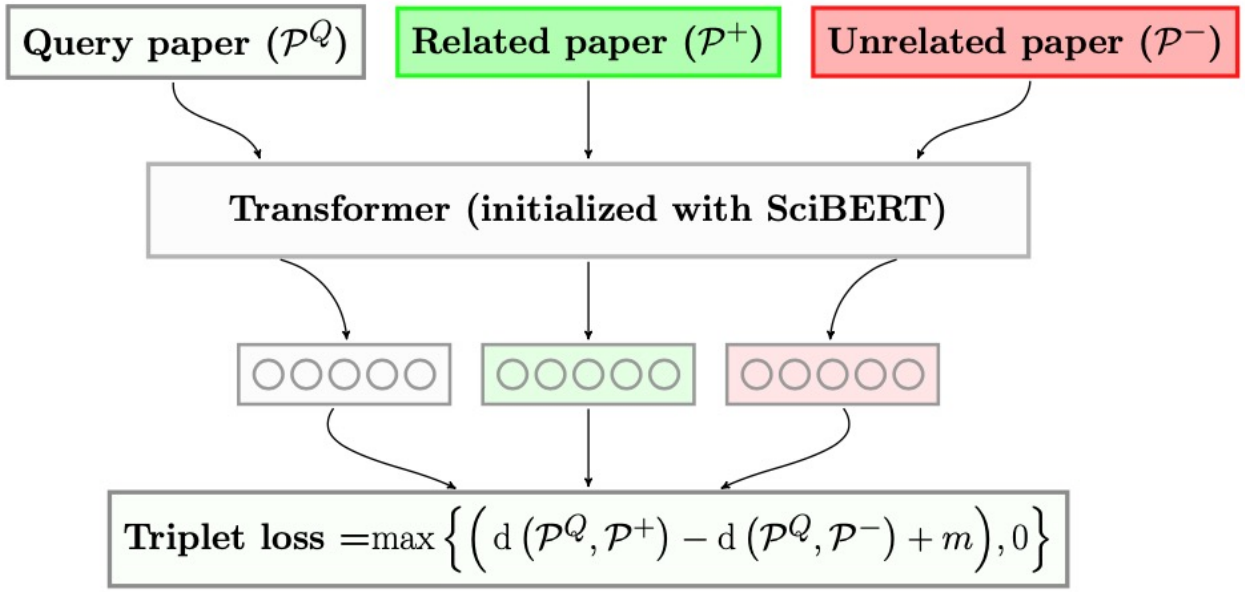


Figure 1: Overview of SPECTER.

SciNCL

- SciNCL is also initialized with weights from SciBERT. It uses the citation graph neighborhood to generate samples for contrastive learning. The underlying citation embeddings are trained on the S2ORC citation graph. It doesn't use direct citation to enforce a hard cut-off to similarity. Instead, it uses controlled nearest neighbor sampling over citation graph embeddings for contrastive learning.

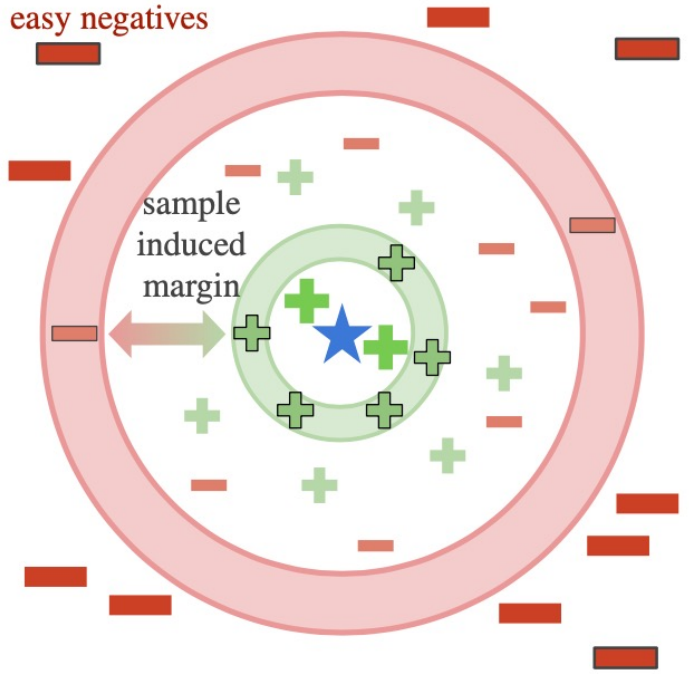
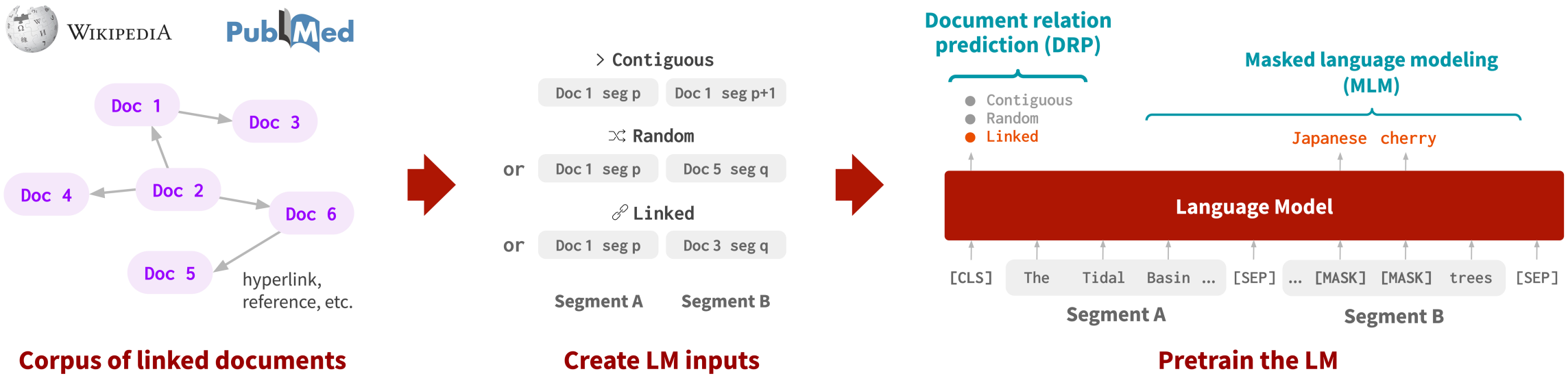


Figure 1: Starting from a query paper ★ in a citation graph embedding space. Hard positives ✚ are citation graph embeddings that are sampled from a similar (close) context of ★, but are not so close that their gradients collapse easily. Hard (to classify) negatives ▭ (red band) are close to positives (green band) up to a *sampling induced margin*. Easy negatives ▬ are very dissimilar (distant) from the query paper ★.

LinkBERT

- LinkBERT is a improvement of BERT that captures **document links** such as hyperlinks and citation links to include knowledge that spans across multiple documents. Specifically, it was pretrained by feeding linked documents from Wikipedia or PubMed. During pretraining, the authors train the LM with two input documents a time and two objectives: masked language modeling (MLM),which predicts masked tokens in the input, and document relation prediction(DRP),which classifies the relation of the two text segments in the input(contiguous, random,or linked).



Cenrtality measure

- Indegree

- Harmonic Centrality
- Harmonic Centrality for a node is the sum of the reciprocals of the distances of u from each vertex to the node in the graph.

- ProNE
- ProNE is a network representation learning model. The design of the embedding ensures a relationship between the cosine values within the embedding and the distances between papers within the citation graph.

SciRepEval: A Multi-Format Benchmark for Scientific Document Representations

- It consists of 25 scientific document tasks to train and evaluate scientific document representation models.
- **Classification**: Fields of study, etc.
- **Regression**: Peer-review ratings, citation counts, etc.
- **Proximity (nearest neighbors) retrieval**: Paper-Reviewer Matching, predicting citation relationship, etc.
- **Ad-Hoc Search:** Clickthrough prediction, TREC-COVID, etc.

Ad-Hoc Search

- Search: Find papers clicked by users based on clickthrough events from a scholarly search engine.
- TREC-CoVID: scores from 0-2
- Feeds Title
- What are feeds?
- They consist of papers recommended for a topic based on a user’s library. This data includes information on whether users found the recommendations relevant or not. The data contains 430 feeds which have more than five positive and two negative paper annotations from users.

Proximity (nearest neighbors) retrieval

- Feeds-1: The first paper added to a feed chronologically serves as the query. The next 5 positive user annotations are considered relevant and 5 negative candidates are sampled either from user annotations or randomly.

- Feeds-M: Given K positive papers annotated in a feed (assuming $K > 5$), we use the first $M = K - 5$ as queries. For every query, the positive candidates are sampled from all the papers the user positively annotated after the query paper was added to their feed, and negative candidates are sampled from user annotations or randomly

Proximity (nearest neighbors) retrieval

- S2AND from AllenAI: 1. Check if two papers are written by the same authors with text of papers and other features. 2. Clustering papers of the same author.

- Same Author Detection based on S2AND: given three papers of which two share an author, the goal is to find the matching pair.

- Highly Influential Citations: if A is cited at least 4 times, in the text of B, we consider it to be highly influential for B.

Proximity (nearest neighbors) retrieval

- Paper-Reviewer Matching: judging whether a given paper is relevant to a potential reviewer.
- Mimno & McCallum (2007), with 393 paper-review relevance ratings from a corpus of 148 NeurIPS 2006 papers and 364 reviewers, annotated by nine human experts.
- Liuetal.(2014),an extension of Mimno&McCallum(2007) which adds 766 additional paper-review annotations.
- Zhaoetal.(2022), with 694 paper-reviewer relevance ratings from a corpus of 75 papers and 1833 reviewers from the IEEE ICIP 2016 conference, annotated by 3 human experts.

Proximity (nearest neighbors) retrieval

- All datasets have been annotated on the same 0-3 relevance rating scale.
- Soft decision: 2 and 3 are relevant. Hard decision: only 3 is relevant.
- The candidate reviewers are all researchers, and we embed all the papers written by them using our models. To obtain the model's score for each candidate reviewer, we compute the cosine similarity between the query paper and each of the candidate's papers, and take the mean of the top 3 similarities as the score.

Proximity retrieval from SciDocs

- Direct Citations
- Co-Citations
- Co-Views
- Co-Reads

Regression

- Citation Count: the authors sample a collection of scientific articles published in 2016 from the set of papers in the search dataset. 5 year period has passed for them to collect citations. Each article has at least one citation, and the citation counts are converted to log scale.

- Year of Publication: the authors sample publications from the search dataset with a publication date after the year 2005 and scale the years so that their values are between 0 and 1.
- The labels are scaled by the mean of the labels in citation count for parity.

Regression

- Peer Review Score: The scores come from ICLR conferences from 2017 to 2022. Each reviewer in ICLR assigns a final rating in the range [0-10], and the authors take the mean rating as the label for every paper.

- h-Index of authors: The authors re-use the peer review score dataset and normalize the h-index to lie between [0,1].

- Tweet Mentions: A dataset created by Jain & Singh (2021) containing tweets about Arxiv papers between 2010-19. Predicting the sum of normalized counts of mentions and retweets.

Classification

- MeSH Desciptors: classifying scientic documents into 30 Medical Subject Headings (MeSH) descriptors provided by National Library of Medicine.

  Ex. "Anti-Bacterial Agents”, "Kidney”, “Brain”

- Fields of Study: classifying papers into 23 Fields of Study. The authors assign Labels in the test set manually. Labels in the train set come from publication venues.

- DRSM (Disease Research State Model) (Burns, 2022): classifying Pubmed papers that deal with six specific aspects of rare diseases. The gold data is annotated by in-house experts and used for evaluation, while the silver data is generated by annotation service providers with medical expertise.

- Biomimicry: the authors sample tags for a set of papers in the PeTaL database (Shyam et al., 2019) to create a binary classification dataset with labels indicating whether each paper is about biomimicry.

Classification from SciDocs

- MeSH Classification: classifying scientific papers to 11 top-level disease classes based on their Medical Subject Headings (MeSH).

- Microsoft Academic Graph (MAG) Paper Topic Classification. 19 topics in total.

Task Format	Name	Train + Dev	Test	Eval Metric	Source
			In-Train		
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:** 54,255	MAP	**This work**
	Citation Prediction Triplets	819,836	—	*not used for eval	Cohan et al. (2020)
SRCH	Search	**Q:** 723,343 **P:** 7,233,430	**Q:** 2,585 **P:** 25,850	nDGC	**This work**
			Out-of-Train		
CLF	Biomimicry	—	11,057	Binary F1	Shyam et al. (2019)
	DRSM	—	7,520 S; 955 G	Macro F1	Burns (2022)
RGN	Peer Review Score	—	10,210	Kendall's $\mathcal{T}$	**This work**
	h-Index of Authors	—	8,438	Kendall's $\mathcal{T}$	**This work**
	Tweet Mentions	—	25,655	Kendall's $\mathcal{T}$	Jain & Singh (2021)
PRX	S2AND	—	**X:** 68,968 **Y:** 10,942	B^3 F1	Subramanian et al. (2021)
					Mimno & McCallum (2007)
	Paper-Reviewer Matching	—	**Q:**107 **P:** 1,729	P@5, P@10	Liu et al. (2014)
					Zhao et al. (2022)
	Feeds-1	—	**Q:** 423 **P:** 4,223	MAP	**This work**
	Feeds-M	—	**Q:** 9025 **P:** 87,528	MAP	**This work**
SRCH	Feeds Title	—	**Q:** 424 **P:** 4,233	MAP	**This work**
	TREC-CoVID	—	**Q:** 50 **P:** 69,318	nDCG	Voorhees et al. (2021)
			SciDocs		
CLF	MAG	—	23,540	Macro F1	Cohan et al. (2020)
	MeSH Diseases	—	25,003	Macro F1	Cohan et al. (2020)
PRX	Co-view	—	**Q:** 1,000 **P:** 29,978	MAP, nDCG	Cohan et al. (2020)
	Co-read	—	**Q:** 1,000 **P:** 29,977	MAP, nDCG	Cohan et al. (2020)
	Cite	—	**Q:** 1,000 **P:** 29,928	MAP, nDCG	Cohan et al. (2020)
	Co-cite	—	**Q:** 1,000 **P:** 29,949	MAP, nDCG	Cohan et al. (2020)

Installation

- git clone git@github.com:allenai/scirepeval.git
- cd scirepeval
- conda create -n scirepeval python=3.8
- conda activate scirepeval
- pip install -r requirements.txt
- pip install -r requirements2.txt

Easy to run the evaluation

- python scirepeval.py -m allenai/specter --batch-size 8 (**You need a GPU to produce the embeddings for a model available on HuggingFace with scirepeval.py.** Please set the batch size to be 8 if you are running this on NEU cluster and are producing embeddings because most GPUs on the cluster can not handle batch size larger than 8.)

- You can specify which task to run in the scirepeval_tasks.jsonl.

- embeddings:”{“save:”}” makes the program produce embeddings for tasks listed in the jsonl file while
- embeddings:”{“load:”}” asks the program to load embeddings directly.

- {"name":"SciDocs MAG","type":"classification","data":{"meta":{"name":"allenai/scirepeval","config":"scidocs_mag_mesh"},"test":{"name":"allenai/scirepeval_test","config":"scidocs_mag"}},"embeddings":{"save":"embeddings/scidocs_mag_mesh.jsonl"},"metrics":["f1_macro"]}

- {"name":"Biomimicry","type":"classification","data":{"meta":{"name":"allenai/scirepeval","config":"biomimicry"},"test":{"name":"allenai/scirepeval_test","config":"biomimicry"}},"metrics":["f1"],"**few_shot**":[{"sample_size":64,"iterations":50},{"sample_size":16,"iterations":100}]}

- ~~If the model is not available on huggingface, you can create a json file for your embeddings. Then load the embeddings for your tasks with scirepeval_tasks.jsonl.~~ This may not be correct for some tasks. I am checking this with ProNE embeddings.

- {"name":"SciDocs CoView","type":"proximity","data":{"simple_format":true, "meta":{"name":"allenai/scirepeval","config":"scidocs_view_cite_read"},"test":{"name":"allenai/scirepeval_test","config":"scidocs_view"}},"embeddings":{"load":"embeddings/scidocs_view_cite_read.jsonl"},"metrics":["map","ndcg"]}

S2AND

- S2AND evaluation needs some extra steps to run and is not included in scirepeval_tasks.jsonl

- Please check the instruction here.
- https://github.com/allenai/scirepeval/blob/main/BENCHMARKING.md#:~:text=%3A%2078.85%0A%20%20%20%20%7D%0A%7D-,S2AND%20evaluation,-S2AND%20evaluation%20requires

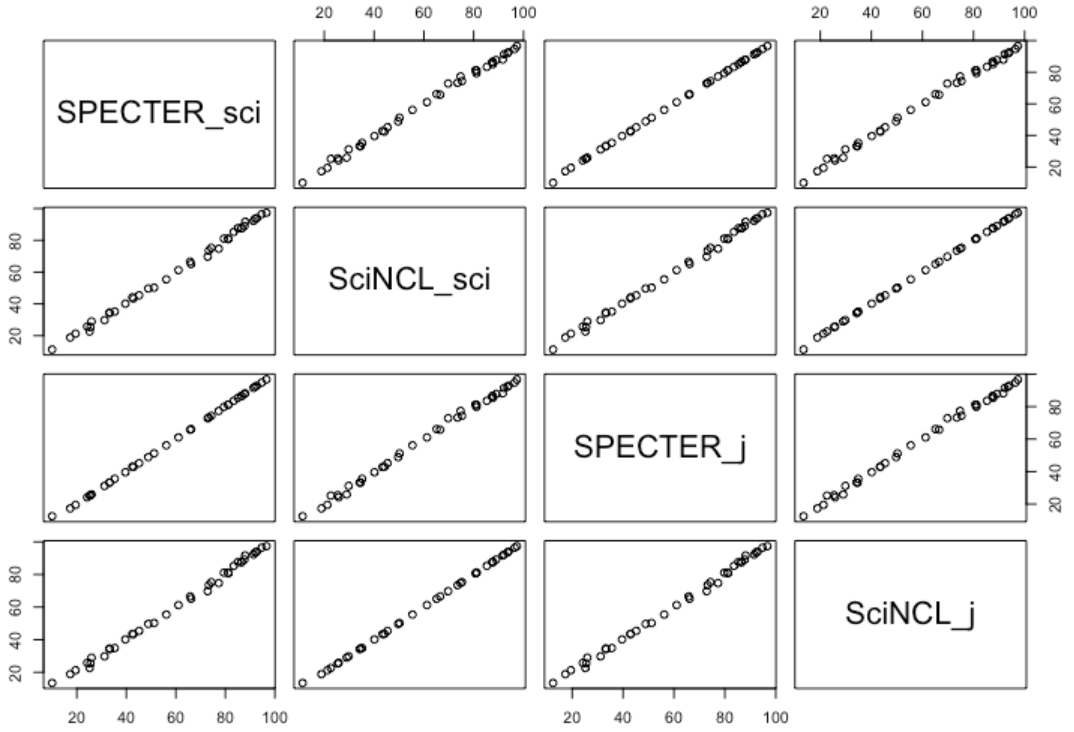
Format of the embedding files

- {"doc_id": "0807.5082", "embedding":…}

- You can find the doc ids from the huggingface datasets: allenai/scirepeval and allenai/scirepeval_test

Scatter plots of the original results and our replication

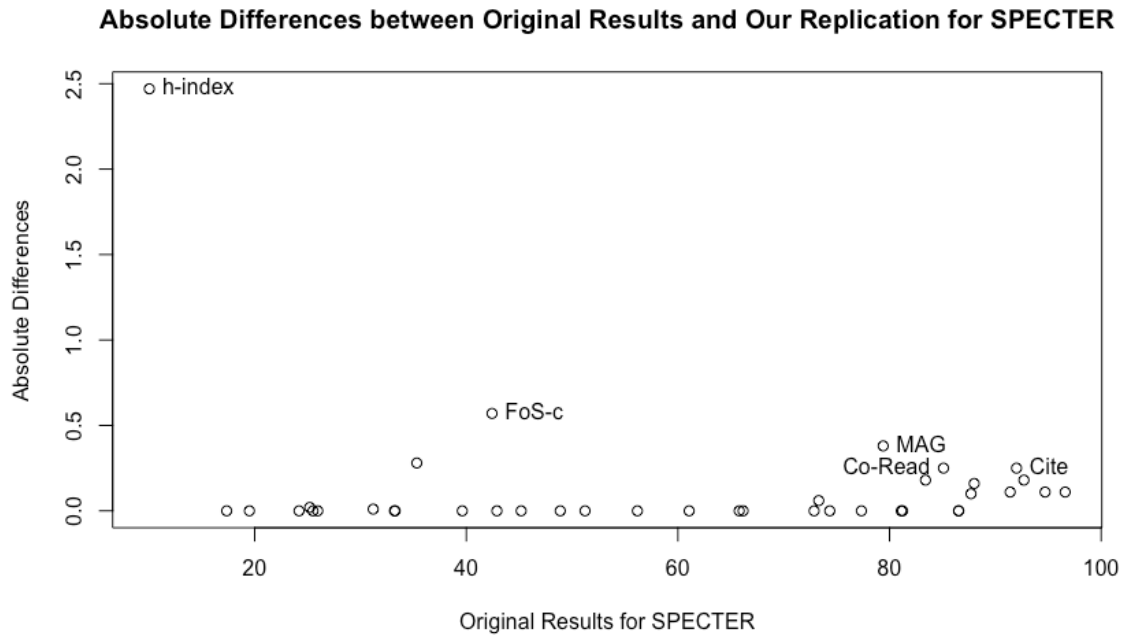
- Github link to more info on reproducing the results.
- https://github.com/allenai/scirepeval/blob/main/BENCHMARKING.md



Correlation between the original results and our replication

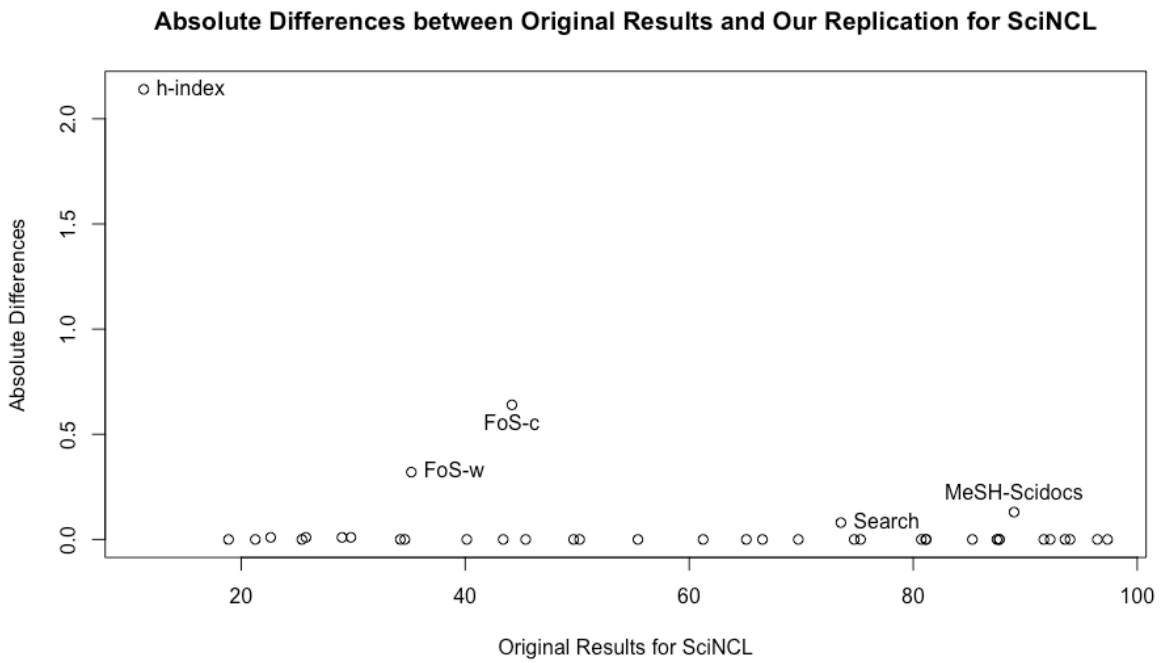
	SPECTER_sci	SciNCL_sci	SPECTER_j	SciNCL_j
SPECTER_sci	1.000	0.999	1.000	0.998
SciNCL_sci	0.999	1.000	0.999	1.000
SPECTER_j	1.000	0.999	1.000	0.999
SciNCL_j	0.998	1.000	0.999	1.000

5 tasks with largest difference for SPECTER



Task	SPECTER_orig	Absolute difference
h-index	10	2.47
Fields of Study-complete	42.4	0.57
MAG	79.4	0.38
Co-Read	85.1	0.25
Cite	92	0.25

5 tasks with largest difference for SciNCL



Task	SPECTER_orig	Absolute difference
h-index	11.3	2.14
Fields of Study-complete	44.2	0.64
Fields of Study-weighted	35.2	0.32
MeSH-Scidocs	89	0.13
Search	73.5	0.08