Reproducibility of benchmarks for scientific document representation models

Presented By:
Shridhar Pawar

Mentor:
Prof. Mayank Singh
Discipline of Computer Science Engineering, IIT Gandhinagar

INTRODUCTION

- We will test three benchmarks on leading models like Bert, SciBERT, and Longformer.
- Our experimentation follows the flow shown in Fig. 1
- We will look into the resourcefulness of benchmark, code reusability, time taken, bugs faced, and diversity of datasets.

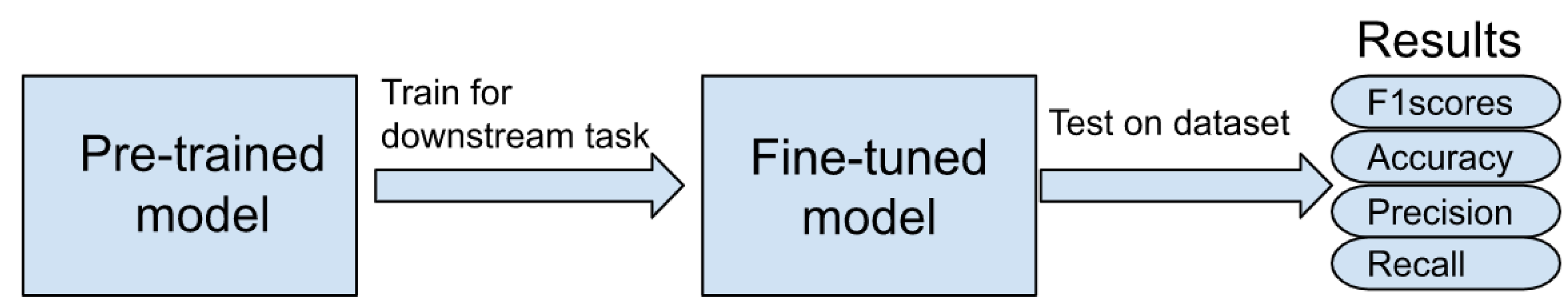

Figure 1. Benchmark Evaluation flow

OBJECTIVE

- We aim to provide insights into benchmarks reproducibility and usefulness in evaluating scientific document representation models.
- This will enable researchers to select the most appropriate benchmark for their specific needs and ensure their results are reliable and reproducible.

METHODOLOGY

1. Run Benchmarks on Bert, SciBert, and Longformer
2. Do error quantification and time taken while running them
3. Analyze the number of tasks, code availability/readability, code documentation

RESULTS/FINDINGS

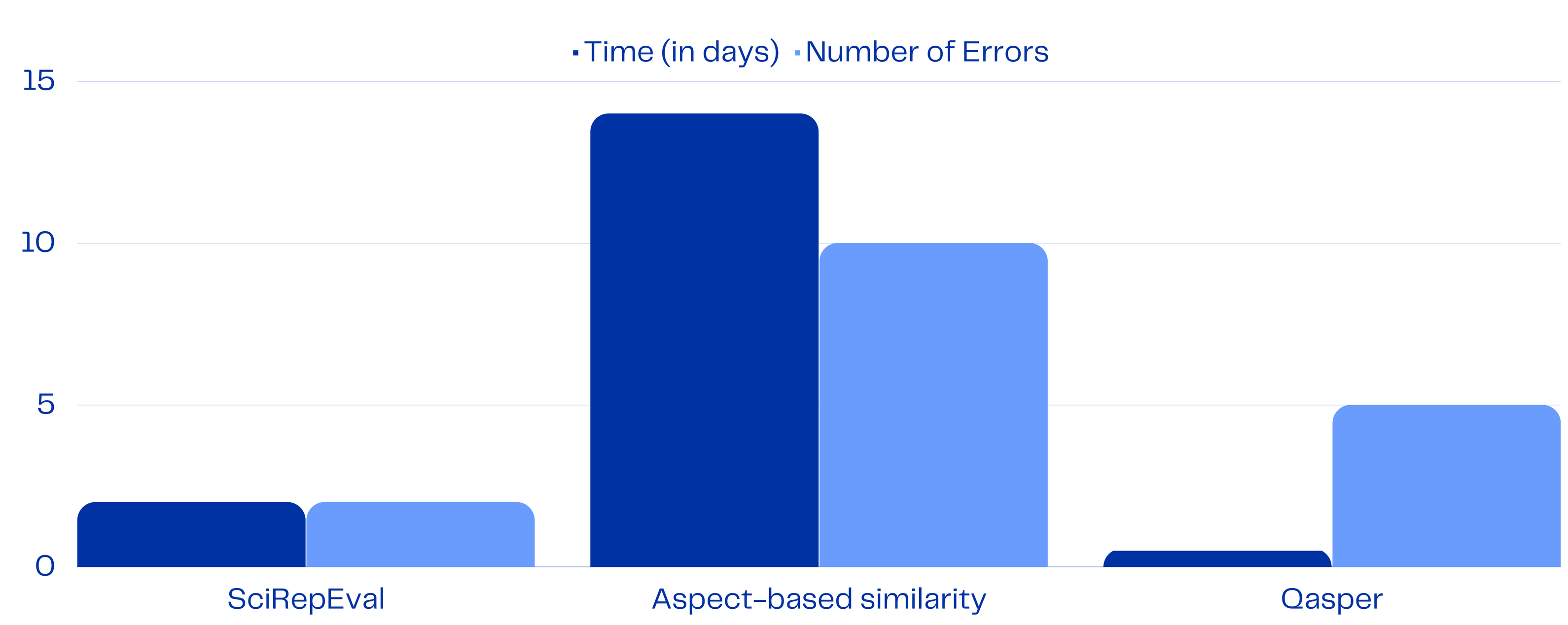

Chart. Comparison of time taken to evaluate, number of errors in the code before successful run between these three benchmarks

	Bert	SciBERT	Longformer
SciRepEval[2]	✓	✓	✓
Aspect-based similarity[3]	Still running ...	Still running ...	Still running ...
Qasper[1]	✗	✗	✓

Table. Results of whether we were able to run these models on benchmarks or not

1. Aspect-based similarity benchmark is not optimized for GPU. It takes weeks to evaluate a simple model such as Bert.

2. Qasper benchmark consists of longer inputs than the typical 512 or 1024 token limit of most BERTlike models. So we can only evaluate models that take long inputs.

3. Qasper benchmark is coded for encoder-decoder type models.

4. SciRepEval has only small errors in creating directories and batch sizes.

5. Debugging in aspect-based similarity benchmark is difficult due to their implementation of model training, which takes weeks before showing an error.

ABOUT BENCHMARKS

1. **SciRepEval :**

- A benchmark suite of 25 tasks across four formats for training and evaluating multi-task embeddings of scholarly papers.
- Included previous benchmark SciDoc as a subset and introduced 11 new tasks, out of which six are explicitly for training.

Task Format	Name	Train + Dev	Test	Eval Metric
		In-Train		
CLF	MeSH Descriptors	2,328,179	258,687	Macro F1
	Fields of study (FoS)	676,524 S	471 G	Macro F1
RGN	Citation count	202,774	30,058	Kendall's $\mathcal{T}$
	Year of Publication	218,864	30,000	Kendall's $\mathcal{T}$
PRX	Same Author Detection	**Q:** 76,489 **P:** 673,170	**Q:** 13,585 **P:** 123,430	MAP
	Highly Influential Citations	**Q:** 65,982 **P:** 2,004,688	**Q:** 1,199 **P:** 54,255	MAP
	Citation Prediction Triplets	819,836	—	*not used for eval
SRCH	Search	**Q:** 723,343 **P:** 7,233,430	**Q:** 2,585 **P:** 25,850	nDGC
		Out-of-Train		
CLF	Biomimicry	—	11,057	Binary F1
	DRSM	—	7,520 S; 955 G	Macro F1
RGN	Peer Review Score	—	10,210	Kendall's $\mathcal{T}$
	h-Index of Authors	—	8,438	Kendall's $\mathcal{T}$
	Tweet Mentions	—	25,655	Kendall's $\mathcal{T}$
PRX	S2AND	—	**X:** 68,968 **Y:** 10,942	B^3 F1
	Paper-Reviewer Matching	—	**Q:**107 **P:** 1,729	P@5, P@10
	Feeds-1	—	**Q:** 423 **P:** 4,223	MAP
	Feeds-M	—	**Q:** 9025 **P:** 87,528	MAP
SRCH	Feeds Title	—	**Q:** 424 **P:** 4,233	MAP
	TREC-CoVID	—	**Q:** 50 **P:** 69,318	nDCG
		SciDocs		
CLF	MAG	—	23,540	Macro F1
	MeSH Diseases	—	25,003	Macro F1
PRX	Co-view	—	**Q:** 1,000 **P:** 29,978	MAP, nDCG
	Co-read	—	**Q:** 1,000 **P:** 29,977	MAP, nDCG
	Cite	—	**Q:** 1,000 **P:** 29,928	MAP, nDCG
	Co-cite	—	**Q:** 1,000 **P:** 29,949	MAP, nDCG

Figure2. SciRepeval benchmark

2. **Qasper:**

- A dataset of 5,049 questions over 1,585 Natural Language Processing papers.

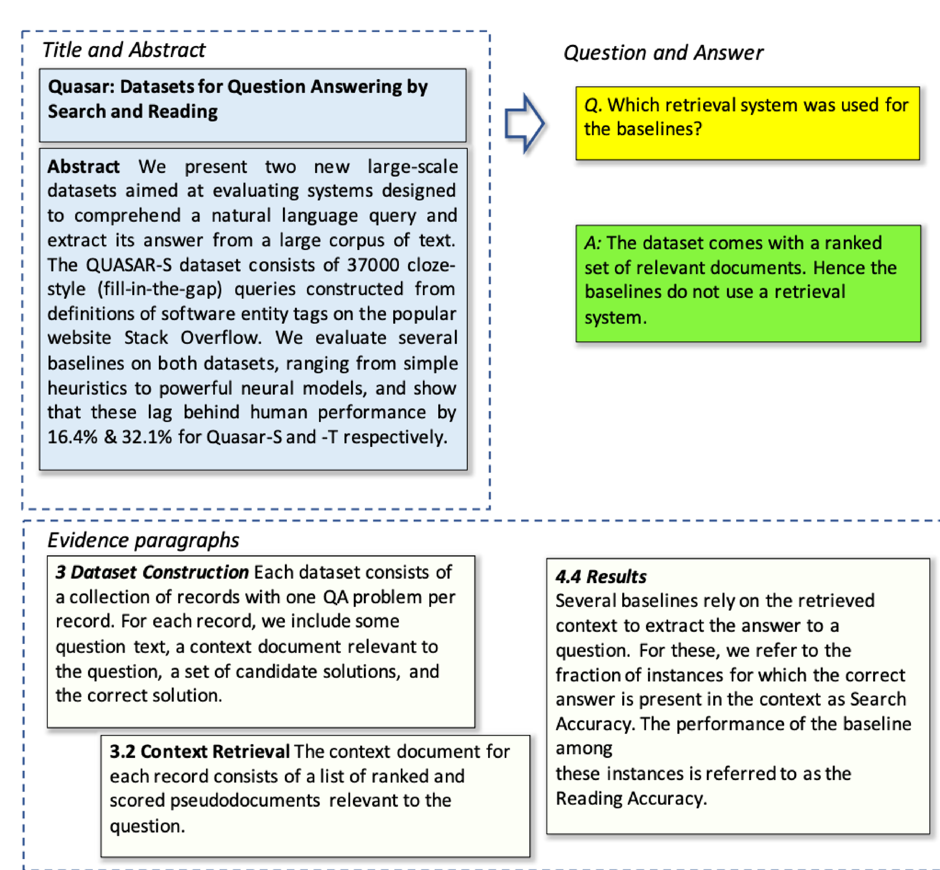

Figure3. Qasper sample question

3. **Aspect-based Similarity:**

- A dataset of 157,606 unique papers with three aspect labels, A = {Task, method, dataset}
- The similarity of documents is computed as the cosine similarity of their vectors.

Aspect	Papers	Labels	Avg. papers per label
Task	154,350	1,421	17.9
Method	108,687	788	12.4
Dataset	37,604	1,743	5.6

Figure4. Number of Labels for each aspect

CONCLUSIONS

- Even though there is much research in scientific document representations, there must be better evaluation benchmarks.
- Except for SciRepEval, other benchmarks do not have multiple downstream tasks.
- Benchmarks need to be better documented and maintained. Also, researchers have to write their codes for evaluating their models. Code reusability needs to be included.

REFERENCES

1. Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N. A., & Gardner, M. (2021). A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. ArXiv. /abs/2105.03011
2. Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2022). SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. ArXiv, abs/2211.13308.
3. Ostendorff, M., Blume, T., Ruas, T., Gipp, B., & Rehm, G. (2022). Specialized Document Embeddings for Aspect-based Similarity of Research Papers. ArXiv. /abs/2203.14541