Reproducibility Of Benchmarks For Scientific Document Representation Models

CS 499: Project Course

Under the guidance of

Prof. Mayank Singh

SUBMITTED BY

Shridhar Sominath Pawar

(19110101)

DEPARTMENT OF COMPUTER SCIENCE

Indian Institute of Technology (Gandhinagar)

Palaj, Gandhinagar (Gujarat) - 382355

January-April 2023

Table of Contents

1) INTRODUCTION

Learning representations of whole documents is critical for various NLP tasks, including classification, search, and recommendation. Recent work has shown how pre-trained language models can be tailored to produce high-quality representations of documents with contrastive learning. While significant progress has been made in evaluating the generalizability of NLP models, the evaluation of scientific document representation models has remained limited. There are very few benchmarks available for evaluating scientific representation models.

In this project, We are studying recently published three benchmarks in this project in this project: SciRepEval, Qasper, and Aspect-based similarity dataset. We are trying to run these benchmarks on leading generalizable deep learning models like Bert, SciBERT, and Longformer. We will look into the resourcefulness of benchmarks, code reusability, time taken, bugs faced, and diversity of datasets. Our experimentation follows the flow shown in Fig. 1

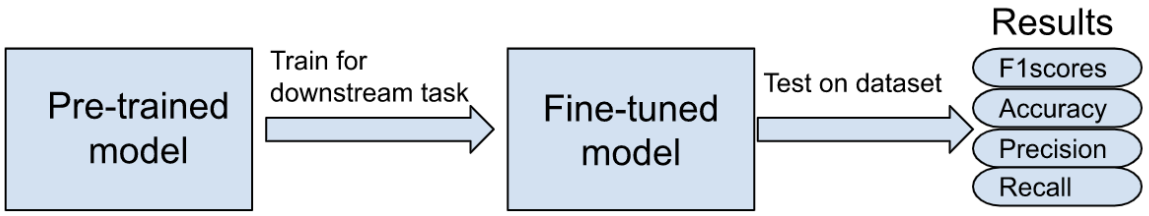

Figure 1. Benchmark Evaluation flow

2) OBJECTIVE

We aim to provide insights into the benchmark's 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.

3) METHODOLOGY

1. Get the publicly available benchmark codes. Try to Run Benchmarks on Bert, SciBERT, and Longformer.

2. We chose Bert because most of the model architectures are similar to Bert in recent years. SciBERT is finetuned Bert model on scientific data. Longformer can take longer inputs (more than 512 or 1024 tokens), making it perfect for document type of inputs.

3. These three combined almost all models available right now, and experimenting with these will give us a generalization of benchmarks on all models.

4. Debug it till the successful run of benchmarks on models and do error quantification and time taken while running them.

5. Analyze the number of tasks, code availability/readability, and code documentation of each Benchmark.

4) ABOUT BENCHMARKS

a) SciRepEval

It is a benchmark suite of 25 tasks across four formats for training and evaluating multi-task embeddings of scholarly papers. SciRepEval aims to enable a comprehensive evaluation of paper embeddings by providing the following:

1. A highly diverse set of tasks spanning multiple task formats such as classification, regression, proximity, and ad-hoc search to challenge the general-purpose applicability of embeddings
2. Realistic tasks that reflect practical use cases of paper embeddings
3. A standard set of both training and evaluation datasets to simplify comparisons between methods evaluated on the benchmark

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

SciRepEval included the previous benchmark SciDoc as a subset and introduced 11 new tasks, out of which six are explicitly for training. More about this dataset can be found here.

b) Aspect-Based Similarity

A benchmark with a dataset of 157,606 unique papers with three aspect labels, A = {Task, method, dataset}. These papers are collected from papers with code websites hosting hand-curated papers in the machine learning domain. Each research paper is labeled with the task the paper is focusing on, the paper's method, and the dataset used. A single paper can have multiple labels for each of these three aspects.

Following are some examples of possible labels:

- Tasks: Low-Rank Matrix Completion, Q-Learning, Quantization, Speaker Recognition, Object Detection
- Methods: Residual Connection, Tanh Activation, Multi-Head Attention, LSTM, Transformer
- Datasets: Atari 2600 Atlantis, Cityscapes, SOP, MS MARCO, Labeled Faces in the Wild

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

Figure3. Number of Labels for each aspect

For each paper, three vector representations are generated for three aspects. The similarity of documents is computed as the cosine similarity of their vectors. More about this dataset can be found here.

c) Qasper

A benchmark of 5,049 questions over 1,585 Natural Language Processing papers. Each question is written as a follow-up to the title and abstract of a particular paper. The answer, if present, is identified in the rest of the paper, along with the evidence required to arrive at it.

To the best of our knowledge, QASPER is the first QA dataset in the academic research domain focusing on entire papers and not just abstracts. More about this dataset can be found here.

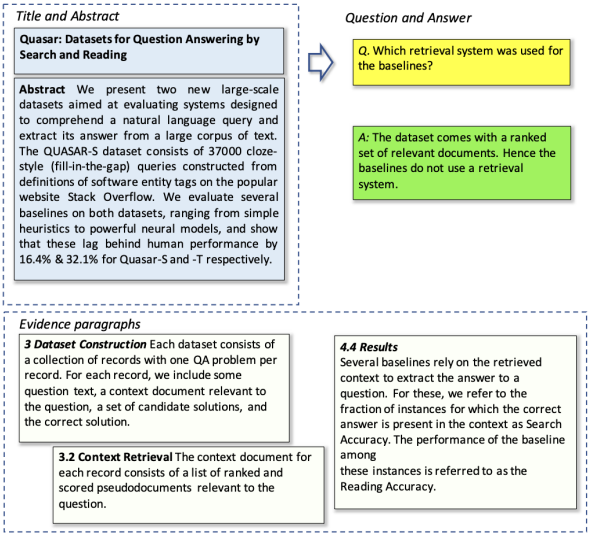

Figure4. Qasper sample question

5) RESULTS

We tried to debug all three benchmarks till we can run three models (Bert, SciBERT, and Longformer) successfully. However Note that we were not able to debug Aspect-based similarity benchmark completely. For Aspect-based similarity benchmark, Time and error shown in following chart are till models are trained.

All experiments were done on single NVIDIA GPU with 32 GB memory.

Time Analysis:

Qasper is quickest to train and evaluate model with 12-13 hours followed by SciRepEval with time around 24-26 hours. Aspect-based similarity benchmark is slowest which runs for weeks.

Chart1. Comparison of time taken to evaluate, number of errors in the code before successful run between these three benchmarks (for Lognfoemer model)

Error Analysis:

- SciRepEval has minimum errors. We have to create some directories manually or change their code to produce directories before running functions and commands from code. Other from that it runs smoothly for any type of model that generates embeddings.

- Qasper has some errors slightly more than SciRepEval. Errors were related to changing default settting of parameters such as batch-size, model, and loading datasets. Note that we can not run Bert and SciBERT on Qasper benchmark (will be discussed below.)

- Aspect embedding benchmark code is not optimized for GPU utilization. It completely runs on CPU and takes weeks to complete training of model before evaluating a model. It had errors in evaluating models. So We had to wait weeks before debugging next error. This made us go back and forth for weeks. Eventually we stopped doing it and concluded that this code is very inefficient.

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

Table1. Results of whether we were able to run these models on benchmarks or not

- SciRepEval runs on all three model. We were not able to run Aspect-based similarity benchmark on any model

- Qasper requires an encoder-decoder Transformer model where the encoder reads the question and the document, and the decoder generates the answer text, which means typical encoder based moels like BERT or decoder based models like GPT are not sufficient.

- Research papers are much longer than the typical 512 or 1024 token limit of most BERTlike models, so only Longformer works for Qasper.

6) FUTURE RESEARCH

Our next goal is to use this debugged code to evaluate our very own model we are developing in the filed of scientific document representation model. Our whole study started when we were searching for right benchmarks for our model.

In future, We can also provide a interface where researchers can choose a benchmark and describe their model and we will provide code with modification for their use.

Also We plan to make our very own benchmark which includes variety of tasks. We aim to provide better code in terms of readability and well documentation.

7) CONCLUSIONS

- Even though there is much research in scientific document representations, there must be better evaluating benchmarks with code reusability.

- Except for SciRepEval, other benchmarks do not have multiple downstream tasks. All are focussed on one task.

- Benchmarks need to be better documented and maintained. Researchers have to write their codes for evaluating their models. Code reusability needs to be included.

8) REFERENCES

1. https://github.com/allenai/qasper-led-baseline
2. https://github.com/allenai/scirepeval
3. https://github.com/malteos/aspect-document-embeddings