A Citations in medical research

Medical research communication follows a standardised format and occurs primarily through publication. A research component is a literature review that contains established facts about a research topic. These established facts are corroborated with references to their source of evidence. References adhere to a standard format and are typically listed at the end of the research article.

It follows, then, that general information about that research can be elicited from the analysis of these references.

A.1 Improvements in encoder performance using citations

Medical CAL TAR has already benefited from citation utilisation. To date, the most performant encoder model ($BioLinkBERT_{base}$) on the CLEF dataset in a CAL setting leveraged citations between research to achieve state of the art. The LinkBERT approach was to view a pertaining corpus as a graph of documents, with each document being a vertex and hyperlinks forming edges between documents. These linked documents were then placed within the same context, which was different from that of BERT random document allocation. A domain-specific variant, *BioLinkBERT*, was created, which pretrained only on PubMed articles (the vertexes) and their citation links (edges). Models were then trained using standard masked language modelling and next-sentence prediction. The performance of a base model (100M parameters) and a large model (340M parameters) was compared to PubMedBert in BLURB, MedQA-USMLE, and MMLU-professional medicine (medical-specific downstream benchmark tasks). $BioLinkBERT_{large}$ achieved state-of-the-art on all reported benchmarks, with an improvement in the BLURB score of 3.2% above PubMedBERT. In the previously reported Goldilock Reproduce study, $BioLinkBERT_{base}$ was used.

This author recreated their experiment with $BioBERT_{large}$ as a classifier model and achieved higher performance in R-Precision in 7 of 12 datasets/policy combinations. The Friedman test for individual datasets found significant differences between the FPT epochs 4 out of 12 times, however, when considering all datasets together, there was no significant difference between the FPT epochs and R-precision for relancy selection policy or uncertainty selection policy. This indicates that the "Goldilocks problem" is not apparent within the large BiolinkBERT model for the CLEF dataset, and that further pretraining does not produce an statistically significant improvement in R-precision. The average R-precision of each FPT epoch is reported in Table 2, with the highest R-precision for relevancy selection policy being 0.847 at FPT ep2 and the highest R-precision for uncertainty being 0.832 at FPT ep1.

Key findings from this research is that an optimal epoch pretraining is unlikely to be found within the CLEF dataset and hence not a viable avenue for future research, however utilisation of citations was beneificial to the CAL process. In terms of experimental design, certain hyperparameters were chosen without clear reasoning (such as batch size being 25, fine tuning for 20 epochs and stopping after 501 documents labelled). This limitation is thought to be a barrier to improving the performance of the encoder CAL process within that experimental framework, given that reported R-P values are already close ceiling.

BioLinkBERT represents a generalised approach that combines citation networks with contextual language understanding. Although this allows the model to capture complex semantic relationships between documents, it also introduces potential noise. When linked documents are placed in the same context during pretraining, the model must process all content within those documents - including sections that may be tangential or unrelated to the citing paper's specific reference. This contextual noise could dilute the precision of the more direct and expertly curated relationships that the citations represent. In contrast, simple citation links directly capture intentional scholarly connections made by domain experts without the additional complexity of processing potentially irrelevant contextual information.

Collection	Dataset size	Model	R-Precision (↑)		Friedman (p)	
			Rel.	**Unc.**	**Rel.**	**Unc.**
Clef 2019 dta test	8	BiolinkBert-Base-ep0	**0.909**	**0.857**	—	
		BiolinkBert-Large-ep0	0.897	0.803	0.914	0.632
		BiolinkBert-Large-ep1	0.827	0.832		
		BiolinkBert-Large-ep2	0.812	0.774		
		BiolinkBert-Large-ep5	0.841	0.814		
		BiolinkBert-Large-ep10	0.881	0.846		
Clef 2017 test	30	BiolinkBert-Base-ep0	0.812	0.794	—	
		BiolinkBert-Large-ep0	0.828	0.797	**<0.05**	**<0.05**
		BiolinkBert-Large-ep1	0.826	**0.827**		
		BiolinkBert-Large-ep2	**0.858**	0.804		
		BiolinkBert-Large-ep5	0.827	0.777		
		BiolinkBert-Large-ep10	0.799	0.757		
Clef 2017 train	20	BiolinkBert-Base-ep0	**0.838**	0.761	—	
		BiolinkBert-Large-ep0	0.778	0.765	**<0.05**	0.28
		BiolinkBert-Large-ep1	0.808	0.789		
		BiolinkBert-Large-ep2	0.767	0.701		
		BiolinkBert-Large-ep5	0.816	0.786		
		BiolinkBert-Large-ep10	0.827	**0.796**		
Clef 2018 test	30	BiolinkBert-Base-ep0	0.794	0.780	—	
		BiolinkBert-Large-ep0	0.789	0.774	0.52	0.50
		BiolinkBert-Large-ep1	**0.812**	0.790		
		BiolinkBert-Large-ep2	0.797	**0.791**		
		BiolinkBert-Large-ep5	0.763	0.773		
		BiolinkBert-Large-ep10	0.763	0.769		
Clef 2019 DTA int. train	20	BiolinkBert-Base-ep0	0.939	0.923	—	
		BiolinkBert-Large-ep0	0.939	0.902	0.78	0.50
		BiolinkBert-Large-ep1	0.941	0.935		
		BiolinkBert-Large-ep2	0.948	0.921		
		BiolinkBert-Large-ep5	0.952	0.945		
		BiolinkBert-Large-ep10	**0.945**	**0.947**		
Clef 2019 DTA int. test	20	BiolinkBert-Base-ep0	**0.934**	**0.900**	—	
		BiolinkBert-Large-ep0	0.899	0.856	0.87	**<0.05**
		BiolinkBert-Large-ep1	0.904	0.840		
		BiolinkBert-Large-ep2	0.909	0.878		
		BiolinkBert-Large-ep5	0.882	0.835		
		BiolinkBert-Large-ep10	0.865	0.841		

Table 1: Performance comparison across different collections and models

Table 2: Average R-precision of each FPT epoch for CLEF dataset

Policy	**ep0**	**ep1**	**ep2**	**ep5**	**ep10**
Uncertainty	0.813	0.832	0.813	0.815	0.814
Relevancy	0.840	0.845	0.847	0.842	0.835

A.2 Citation network mining within medicine research

Simpler approaches to citation network mining exist within the target domain itself, as it has long been established that references contain.