Goldilocks V3

Aaron HA Fletcher

September 2024

1 Introduction

High-recall retrieval (HRR) tasks involve identifying all or most documents of interest within a large collection. HRR tasks are suitable where the need for high recall outweighs the need for precision, such as in cases of systematic review in medicine, electronic document discovery in Law or online content moderation. Technology-assisted reviews (TAR) are automated methods that reduce the total number of documents reviewed in HRR tasks.

- SR in medicine, outline the process.
- Motivations for using TAR for HRR tasks.
- Motivation for using LLMs in TAR.

  This paper adds to the current research on CAL TAR process by exploring if the "Goldilocks problem" exists within BiolinkBERT model.

Flesh out with information regarding SRs

2 Background

A successful approach to TAR is a continuous active learning approach (CAL) in which an oracle, such as a human, is iteratively queried with documents from an unlabelled total pool (starting with a very small number). At each iteration, this labelled pool is then used to incrementally train a classifier to rank documents, and further documents are selected for Oracle labelling. Once the stopping criteria have been met, the aim is to have most of the relevant documents labelled or ranked highly.

There have been multiple approaches to using CAL for how, however, work within CAL for TAR can be delineated by the type of models used within the CAL process and how the document is represented within this process. Early work in CAL followed the AUTOTAR approach, in which a feature-based classifier, such as SVM or linear regression, which was paired with a saturated TF-IDF document representation, which is broadly outlined in Algorithm 1. Despite being computationally inexpensive and achieving moderate sucess in recall, these approaches suffered primarily from limitations within the document

representation, which was unable to capture the semantic meaning between terms within the document, ignoring long range dependencies and positional information. With the development of transformer architecture, large language pre-trained models (LLMs) also found success within the CAL process.

Algorithm 1 The AUTOTAR CAL approach

1: Find a relevant "seed" document using ad hoc search, or construct a synthetic relevant document from the topic description
2: Initialize training set with the seed document from step 1, labeled "relevant"
3: Set initial batch size $B \leftarrow 1$
4: **repeat**
5: Temporarily add 100 random documents from the collection to the training set, labeled "not relevant"
6: Train an SVM classifier using the training set
7: Remove the temporary random documents added in step 5
8: Select the B highest-scoring documents for review
9: Review the documents, coding each as "relevant" or "not relevant"
10: Add the reviewed documents to the training set
11: $B \leftarrow B + B/10$
12: **until** sufficient number of relevant documents have been reviewed

Adapting LLMs on downstream tasks utilises transfer learning. This is where a model is pretrained using a pretraining objective, such as masked-language modelling (MLM) and then that model is adapted by being fine-tuned on a downstream task. The process of pre-training a large language model can be succinctly defined as training a model using a corpora and minimising loss generated from a pre-training objective (i.e. MLM or next sentence prediction). This pre-trained model is typically adapted through using more domain adaptation (further pretraining (FPT) on a task-specific corpora) and/or through changing the pre-training objective, such as in the CAL process to binary classification.

Early encoder-only architectures, such as BERT, offered an incremental improvement for the CAL process compared with the AUTOTAR approach, albeit with a trade-off in computational cost. In this approach, the feature-based classifier was replaced with the BERT model, which was FPT on a task-specific corpora, and fine-tuned for classification. This approach represents documents through a self-attention mechanism, capturing semantic meaning between terms and long-range dependencies. The general approach is outlined in Algorithm 2.

Initially, adapting BERT to the TAR process was difficult, often with it performing on par with the AUTOTAR approach Continuous active learning using Pretrained transformers. It was found that BERT-based TAR performed better being FPT on the unlabelled collection effectively adapting pretrained lm for AL and with a specific number of FPT epochs, with the optimal number being dependent on the dataset, aka the "Goldilocks problem". It was termed Goldilocks because FPT was necessary, yet excessive FPT can occur. Interestingly, this "Goldilocks" stopping further pretraining epoch was not the same for all datasets, and not even consistent across reproducibility studies within the

Algorithm 2 The BERT-based CAL approach

1: Fine-tune BERT language model on the full document collection using MLM
2: Find a single random relevant document as the seed
3: Initialize training set with the seed document from step 2, labeled "relevant"
4: **repeat**
5: Fine-tune BERT for classification using all labeled documents
6: Use fine-tuned BERT to score all unlabeled documents
7: Select 25 documents using active learning strategy (relevance feedback or uncertainty sampling)
8: Review the selected documents, labeling each as "relevant" or "not relevant"
9: Add the newly labeled documents to the training set
10: **until** 20 iterations completed or other stopping criterion met
11: Use final fine-tuned BERT model to rank remaining unlabeled documents

same dataset. In some datasets, such as CLEF, it didn't exist. Explanations for why this might be the case was not explored, however, it could be suggested that FPT instability could be a contributing factor.

Consistent findings across research using BERT for CAL was that the pretraining domain corpora aligning with the downstream task for the recall task was important to producing better classifiers. Indeed, the Reproducability paper found that using the pretrained BioLinkBERT over BioBERT resulted in better CAL performance on the CLEF dataset. BioBERT's is initialised from existing BERT weights (which copora contained English Wikipedia and BookCorpus), and then further pre-trained and Pubmed Abstracts and PMC full-text articles. BiolinkBERT in contrast is initialised from PubMeBert existing weights (which pretraining corpora contained only PubMed abstracts and PMC full-text articles) and leverages hyperlink information to improve the semantic meaning of the model. A salient difference between the two approaches (FPT from BERT existing weights vs FPT from domain-specific corpora weights) is that the vocabulary of FPT from domain-specific corpora matches more closely with the downstream task vocabulary, which is not the case for FPT from BERT existing weights.

The Reproducability paper's main relevant findings to this work were:

- Goldilocks problem wasn't apparent within BioBERT model on the CLEF dataset.
- Using a more domained aligned pretrained model (BioLinkBERT) resulted in superior performance to a less aligned pretrained model (BioBERT) with domain-specific FPT.

The work did not explore if the "Goldilocks problem" exists within these more domain aligned pretrained models, or to put it another way, if superior performance could be achieved within these domain aligned models through FPT. This work aims to address this gap in research.

This is the research gap from the previous work

LLMs with different pretraining objectives have been found to perform better in classification tasks. So far, work has utilised MLM, where the typically 15% of tokens are masked during pre-training, with the model being trained to predict these masked tokens. Recent work, ELECTRA, has explored using a corrupted token objective, where the model is trained to predict if tokens within a sequence are corrupted or not. Corrupted token prediction (CTP) pre-training objective outperforms MLM in classification tasks. To date, no work has utilised CTP in the CAL process. .

This is the section half of the paper, and the most novel part of the work

3 Datasets

This research focuses on CLEF datasets and Synergy Dataset. .

4 Research Questions

1. *Is the "Goldilocks problem" apparent on the BioLinkBERT model using the CLEF dataset?*
2. *Does using a model trained using a corrupted token pretraining objective result in superior CAL performance than the masked-language modelling objected on the CLEF dataset?*
3. *Are there features within the CLEF dataset that correlate with performance?*

Mention about features which make BERT CAL performance better

No one has used the Synergy Dataset for TAR before

5 Experimental Setup

6 Hardware and Hyperparameters

All experiments were conducted on the University of Sheffield's High Performance Computing cluster. A100 GPUs with 80GB of memory were used for all experiments using pretrined LLM models, and baseline experiments were conducted on a single CPU. Hyperparameters were taken from the Reproducability paper

- Batch size: 25
- Further pretraining on unlabelled collection.
- Total of 20 iterations of active learning loop (501 total oracle reveals)
- Total of 20 fine-tune epochs per active learning loop.

7 Evaluation Metrics

Recall-precision (R-precision) was reported for all experiments. R-precision determines the percentage of relevant documents that are ranked in the top k documents, where k is the total number of relevant documents in the collection. R-precision reported is the R-precision as calculated after the final active learning iteration (iteration 20). Within systematic reviews, a R-precision of 0.95 (95%) is considered sufficient.

Unlike the reproducibility paper which used paired t-tests with bonferroni correction per dataset and selection policy, statistical significance of the differences between the FPT epoch and R-precision was calculated using Friedman test over all datasets grouped by selection policy. Since the goal is to find differences between treatment groups (i.e. FPT epochs), the Friedman test is more appropriate for comparing multiple groups simultaneously, and bonferroni correction was not required.

8 Results

8.1 RQ1

Using the BiolinkBERT large model surpassed the performance of the BiolinkBERT base model 7 out of 12 times. 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.

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		
		BiolinkBert-Large-ep1	0.827	0.832		
		BiolinkBert-Large-ep2	0.812	0.774	0.914	0.632
		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		
		BiolinkBert-Large-ep1	0.826	**0.827**		
		BiolinkBert-Large-ep2	**0.858**	0.804	**<0.05**	**<0.05**
		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		
		BiolinkBert-Large-ep1	0.808	0.789		
		BiolinkBert-Large-ep2	0.767	0.701	**<0.05**	0.28
		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		
		BiolinkBert-Large-ep1	**0.812**	0.790		
		BiolinkBert-Large-ep2	0.797	**0.791**	0.52	0.50
		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		
		BiolinkBert-Large-ep1	0.941	0.935		
		BiolinkBert-Large-ep2	0.948	0.921	0.78	0.50
		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		
		BiolinkBert-Large-ep1	0.904	0.840		
		BiolinkBert-Large-ep2	0.909	0.878	0.87	**<0.05**
		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