How does Feedback Signal Quality Impact Effectiveness of Pseudo Relevance Feedback for Passage Retrieval?

Hang Li
The University of Queensland
Brisbane, Australia
hang.li@uq.edu.au

Ahmed Mourad
The University of Queensland
Brisbane, Australia
a.mourad@uq.edu.au

Bevan Koopman
CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Guido Zuccon
The University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

ABSTRACT

Pseudo-Relevance Feedback (PRF) assumes that the top results retrieved by a first-stage ranker are relevant to the original query and uses them to improve the query representation for a second round of retrieval. This assumption however is often not correct: some or even all of the feedback documents may be irrelevant. Indeed, the effectiveness of PRF methods may well depend on the quality of the feedback signal and thus on the effectiveness of the first-stage ranker. This aspect however has received little attention before.

In this paper we control the quality of the feedback signal and measure its impact on a range of PRF methods, including traditional bag-of-words methods (Rocchio), and dense vector-based methods (learnt and not learnt). Our results show the important role the quality of the feedback signal plays on the effectiveness of PRF methods. Importantly, and surprisingly, our analysis reveals that not all PRF methods are the same when dealing with feedback signals of varying quality. These findings are critical to gain a better understanding of the PRF methods and of which and when they should be used, depending on the feedback signal quality, and set the basis for future research in this area.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; **Information retrieval query processing**.

KEYWORDS

Pseudo Relevance Feedback, Feedback quality, Dense retrievers

ACM Reference Format:
Hang Li, Ahmed Mourad, Bevan Koopman, and Guido Zuccon. 2022. How does Feedback Signal Quality Impact Effectiveness of Pseudo Relevance Feedback for Passage Retrieval?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531822

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531822

1 INTRODUCTION AND RELATED WORK

A key assumption in Pseudo-Relevance Feedback is that the top-k documents used as feedback are relevant. Consider for example the scoring formula of the popular Rocchio method [13][1]:

$$\vec{q'} = \alpha\vec{q} + \beta\frac{1}{|Rel|}\sum_{d_i inRel}\vec{d_i} \tag{1}$$

where, the vectors $\vec{d_i}$ refer to the top-k documents retrieved by the original query $\vec{q}$. Similar treatments are employed by other PRF techniques, both those for bag-of-words models [1, 5, 10, 12**?** , 13] and for neural models [6, 7, 18].

This assumption is often incorrect, i.e. the top-k signal often contain irrelevant documents. Then how do PRF methods behave in the presence of different quality of the relevance signal, e.g. if all k documents are relevant vs. if all of them are not relevant? And do PRF methods differ in their behaviour when examining signals of different quality, e.g., a method that is more effective than another when the relevance signal is of high quality, exhibits large losses compared to the other method when the feedback signal is of poor quality? These aspects have often been ignored in the PRF literature, and there is no systematic understanding of how PRF behaves depending on the feedback signal quality, nor how results from methods differ depending on the feedback quality. However, these are important considerations to make, for a number of reasons.

First, PRF has been shown to provide mixed effectiveness [1]. The factors affecting PRF effectiveness may be many, and certainly include representation choices, PRF depth, and method-specific settings (e.g., for Rocchio, these would be the weights α and β) [14]. In addition, we posit that the feedback signal quality also plays a fundamental role in shaping PRF's effectiveness – our empirical results reinforce this standing.

Second, PRF is often studied in the context of a standard first-stage retrieval method, commonly BM25, and statements regarding the comparative effectiveness of different PRF methods are made in this context. However, it is unclear whether these statements would be valid if a stronger, or weaker, first-stage retrieval was used. This is important because, in practice, many would transfer the findings from such research into production systems that may differ in terms of the quality of the feedback signal provided to the PRF technique.

[1] The Rocchio method also has allowance for negative feedback – we removed this here for brevity; we also note that in the PRF setting, negative feedback is often not used.

In this paper we provide the first systematic understanding of how feedback signal quality impacts the effectiveness of PRF techniques. We do this in the context of the Rocchio method for bag-of-words models and of two PRF methods for dense retrievers, techniques that have recently gained momentum both in the research literature [7, 18] and in practical adoption [6].

2 METHOD

The goal of this study is to evaluate how the quality of feedback signals affects the performance of PRF methods. To achieve this, we devise two sets of different experiments to control the PRF signal from two different sources. In this section, we outline how we perform PRF with the controlled feedback signals on top of different initial retrieval methods.

2.1 Controlling the Quality of the Feedback Signal

In all experiments, we consider three different levels of quality of the feedback signal: strong, moderate, and weak. In the experiments we use the MS MARCO passage ranking dataset [11] and the queries from the TREC Deep Learning Track Passage Retrieval Task 2019 [2] (TREC DL 2019) and 2020 [3] (TREC DL 2020); detailed statistics for these datasets are given in Table 1. Assessments in these datasets are graded on a 4-point relevance scale – 0: irrelevant; 1:relevant; 2: highly relevant; 3: perfectly relevant. We define the feedback signal as *strong* if all k passages included in the signal have relevance label 2 or 3. We define the signal as *moderate* if all k passages have label 1; otherwise if all passages have label 0 we define the signal and being *weak*. For simplicity, we do not consider *mixed signals*, where the relevance of the passages in the top k varies, but it is easy to extend this work in that direction. In terms of PRF depth k, we study $k = 1$ and $k = 3$. This choice was made because the ANCE-PRF [18] model checkpoint shared by the original author has been created for $k = 3$ and thus not optimised for higher values of k. We also note that Yu et al. [18] investigated other depths settings from 0 to 5 and found that the checkpoint with $k = 3$ provides the highest effectiveness. Furthermore, we highlight that depths values larger than 5-6 are not possible in ANCE-PRF on the MS MARCO corpus because the text of passages beyond those values would be ignored by ANCE-PRF, due to the limited size of input the ANCE encoder accepts.

We first consider the feedback signal obtained from a first-stage retriever. As first stage retrievers, we consider a representative bag-of-words method, BM25 [12], and three representative dense retrievers methods, namely ANCE [16], TCTv2-HN [9], and DistillBERT-Balanced [4]. Once the initial retrieval is performed (results up to rank position 1,000), we filter the results to remove all unjudged passages. Then, we filter once more to form three distinguished rankings by only considering passages with labels 2 and 3 (for strong signal), 1 (for moderate signal) and 0 (for weak signal). From each set, we then sample 12 passages for each query; if a query does not have 12 passages in one of the three sets (e.g., has less than 12 passages with label 1), then the query is discarded from all sets and ignored for the evaluation. The statistics for the resulting filtered datasets are also reported in Table 1. The rationale for choosing 12 passages is as follows. First, we recorded the number of judged passages for each relevant level for each query. From this distribution we then identified the smallest amount of judged passages across any label – choosing this amount of passages in our experiment would guarantee that every query then has the same amount of unique passages for signal type. However, since the depth k values we experimented with are 1 and 3, we also need to ensure the number of selected passages is a multiple of 1 and 3. This last requirement resulted in identifying 12 as the largest suitable number of passages to select[2].

Table 1: Statistics of the two datasets considered in our experiments. The statistics of the datasets after we remove the queries that do not have enough judged passages are labeled with (Filtered). We use the Filtered datasets in our experiments.

	#Queries	#Passages	Avg #J/Q	#Judgements
TREC DL 2019	43	8,841,823	215.3	9,260
TREC DL 2019 (Filtered)	36	8,841,823	217.7	7,838
TREC DL 2020	54	8,841,823	210.9	11,386
TREC DL 2020 (Filtered)	42	8,841,823	212.8	8,936

When $k = 1$, we use the 12 passages for each query to generate 12 runs using for each a different passage from the set as the relevance feedback signal. Then, the runs for a query are averaged and results are reported. Thus, for each query, we have 3 main results, one for each level of the feedback signal (i.e. strong, moderate, weak) – each of these was obtained by averaging the results obtained from 12 instances of the corresponding signal.

The process when $k = 3$ is similar, apart that, for each query, we randomly split the 12 passages into 4 groups, each containing 3 feedback passages. For each query, we take all 4 groups, and perform PRF, to produce 4 runs for a single query, then we average the performance of these 4 runs to get the final performance of for that query, on a specific level of feedback quality. Thus, for each query, we have 3 main results, one for each level of the feedback signal (i.e. strong, moderate, weak) – each of these was obtained by averaging the results obtained from 4 instances of the corresponding signal and each of these contained 3 passages.

We then repeat the settings above, but sampling passages from the qrels [3] rather than from the baseline runs. We do this to remove any influence of a strong or weak first-stage retrieval on our findings. Queries that were excluded before because the rankings contained less than 12 passages of any given label are also ignored here.

2.2 Considered PRF Methods

While there are many methods of retrieval and PRF being proposed in the literature, in this first investigation we consider a subset of these methods that allows us to understand what the impact of

[2] Note: we removed 7 out of 36 queries from TREC 2019 (removed query ids: '1124210', '443396', '855410', '1117099', '1037798', '1121709', '131843') and 12 out of 42 queries from TREC 2020 (removed query ids: '1116380', '405163', '42255', '1105792', '1115210', '324585', '1131069', '673670', '336901', '768208', '1030303', '258062'). These queries were removed because for each of these queries at least a label was not sufficiently represented (i.e. less than 10 passages).

[3] The file containing the relevance assessments.

feedback quality is on PRF effectiveness with respect to representation type, i.e. bag-of-words vs. dense vectors, and PRF type, i.e. learnt vs. not learnt.

Based on this, we decided to use BM25 as a representative bag-of-words method, noting that differences with other methods such as Language Modelling are often not substantial, along with ANCE [16], TCTv2-HN [9], and DistillBERT-Balanced [4] as representative dense retrievers. Note that ANCE is based on RoBERTA, TCTv2-HN on BERT and DistillBERT-Balanced on a reduced version of BERT (learnt with knowledge distillation), and thus do differ to some extent in terms of representation.

Similarly, we selected the ANCE-PRF method [18] and its extensions to TCTv2-HN and DistillBERT-Balanced by Li et al. [7], as representative *learnt* PRF methods. In these methods, in fact, a PRF encoder is fine-tuned to the relevance feedback task. We note that bag-of-words models do not have a corresponding complex trainable method (often tuning is performed but involves optimizing one or a handful of parameters, not the millions of parameters in the considered transformer-based models). We then selected the Vector-PRF method by Li et al. [6]; specifically we used the Rocchio variant of their method, which follows the general Rocchio PRF formula of Equation 1, but where the vectors are the actual dense representations from the dense encoders used for the first stage retrieval. The parameters in these methods are only two (α, β) and we set them to the values used in previous work [6]. The method can be applied on top of any dense retriever, and we apply it to the 3 dense retrievers considered here. This method has an obvious correspondence in the bag-of-words space: it's the original Rocchio method – thus we consider Rocchio PRF on top of BM25, rather than the more popular RM3 method, to have a direct comparison between bag-of-words and dense retrievers under the same PRF strategy.

For all methods, be it bag-of-words or dense retrievers, learnt PRF (a.k.a. ANCE-PRF and derivatives) or Vector-PRF, we use the implementations available in Anserini/Pyserini [8, 17] and the checkpoints made available by the corresponding authors of the techniques. We implement Rocchio PRF on top of the bag-of-words model in Pyserini and add this implementation to the GitHub repository associated with our paper[4].

3 RESULTS

3.1 Signal Quality and PRF Methods

First, we investigate the interplay between signal quality and the different PRF methods. Table 2 reports MAP, Reciprocal Rank (RR), nDCG@1,3,10,100, Recall@1000 (R@1000) for the effectiveness of each model with different PRF signal qualities. For simplicity, we only show model effectiveness with PRF depth 3, since either PRF depth 1 or 3 show similar trends.

Rocchio PRF. We use Rocchio [13] on top of the bag-of-words retrieval model BM25 as well as an adaptation of the Rocchio method for dense retrievers, called Vector PRF [6]. For the dense retrievers, we applied Vector PRF on top of ANCE, TCTV2-HN and DistillBERT. The parameter settings are presented in Table 3. For BM25, the Rocchio parameters were set to $\alpha = 0.75$ and $\beta = 0.15$, following previous literature. For all dense methods, they were set to $\alpha = 0.6$ and $\beta = 0.4$ on TREC DL 2019 and on 2020 (only when $k = 3$) and to $\alpha = 0.5$ and $\beta = 0.5$ on TREC DL 2020 when $k = 1$. These choices were made based on the results from Li et al. [6].

Although somewhat tuned, then, this Rocchio PRF method was not "learnt" (as opposed to the learnt PRF methods below).

When bag-of-words are used, the PRF signal extracted from the first stage without further filtering (uncontrolled PRF signal) only improves R@1000 with PRF depth 1,3 and nDCG@3 with PRF depth 1 on TREC DL 2019; nDCG@1 is on par with the BM25 baseline on TREC DL 2020; and all other metrics exhibit drops after the use of PRF. However, when we control the quality of the PRF signals, strong signals substantially enhance the effectiveness over all metrics and datasets; moderate signals marginally improve R@1000; weak signals hurt the effectiveness significantly across all metrics, with some losses even larger than 60% compared to the BM25 baseline model.

When dense retrievers are used, the uncontrolled PRF signal gives rise to improvements across the majority of metrics on both datasets. With strong PRF signals, the improvements are significant across all metrics on both datasets, except for TCTV2+VPRF-Rocchio in nDCG@1 on TREC DL 2019 with PRF depth 3. When the PRF signals are moderate, ANCE+VPRF-Rocchio still achieves significant improvements in terms of MAP, R@1000, and nDCG@100; for TCTV2+VPRF-Rocchio, however, effectiveness decreases quickly compared to the strong signals, resulting in most metrics being now significantly worse than the baseline models on both datasets. A similar behaviour occurs for DistilBERT+VPRF-Rocchio. When the weak PRF signals are used, improvements in TREC DL 2019 are observed only for R@1000 with PRF depth 3 for all three models; improvements in TREC DL 2020 are observed only for ANCE+VPRF-Rocchio with PRF depth 3 and TCTV2+VPRF-Rocchio with PRF depths 1 and 3. With DistilBERT+VPRF-Rocchio in TREC DL 2020 all metrics are worse than the baseline and some losses are larger than 40%.

Learnt PRF. We use ANCE-PRF [18] and its variants TCTV2-PRF and DistilBERT-PRF [7] as example of learnt PRF methods on dense representations. Bag-of-words representations do not have an equivalent, heavily learnt PRF method.

When not controlling the quality of PRF signals, all three models substantially improve the respective models without PRF on most metrics across both datasets. When the PRF signals is strong, all three models improve significantly more over the baseline models on both datasets, except TCTV2-PRF for nDCG@1 on TREC DL 2019. By using moderate signals, for all three models larger improvements only occur for deep metrics, such as MAP, nDCG@100, and R@1000. For other metrics instead, effectiveness is either on par or worse than the baseline models (without PRF), on both datasets. For weak signals, marginal improvements can still be observed for deep metrics, but these are much smaller than for other PRF signal qualities, while losses are abundant and some are larger than 20%.

In conclusion, with strong PRF signals, Rocchio PRF approaches, either BM25+Rocchio or Vector PRF, can improve the performance across all metrics on both datasets. However, when we change to use only moderate signals, BM25+Rocchio only can marginally improve deep recall, where ANCE+VPRF-Rocchio is more resilient to this

[4] https://github.com/ielab/Noise-PRF

Table 2: Effectiveness of PRF methods across different representations and PRF signal qualities. *R* stands for the Rocchio PRF method for bag-of-words, baselines are the PRF runs without control of the PRF signal quality (i.e., standard PRF on top *k* retrieved documents). For each signal quality, the PRF models are divided into three categories: Rocchio PRF on bag-of-words, VectorPRF-Rocchio on dense retrievers, and trained PRF on dense retrievers. Statistical significance analysis is performed using two-tailed paired Student's ttest with Bonferroni correction; significant differences are marked with †.

	Models	TREC DL 2019							TREC DL 2020						
		MAP	RR	R@1000	nDCG@1	nDCG@3	nDCG@10	nDCG@100	MAP	RR	R@1000	nDCG@1	nDCG@3	nDCG@10	nDCG@100
BASELINE	BM25	0.2697	0.7044	0.7687	0.5972	0.5298	0.4971	0.4945	0.2870	0.6531	0.7938	0.5595	0.5155	0.4959	0.4959
	ANCE	0.3908	0.8501	0.8031	0.7222	0.7022	0.6767	0.5860	0.4047	0.8275	0.7804	0.7619	0.7479	0.6806	0.5670
	TCTV2-HN+	0.4676	0.8788	0.8794	0.8009	0.7488	0.7309	0.6631	0.4047	0.8275	0.7804	0.7619	0.7479	0.6806	0.5670
	DistilBERT	0.4832	0.8763	0.8905	0.7454	0.7466	0.7319	0.6698	0.4742	0.8677	0.8770	0.7778	0.7887	0.7207	0.6382
	BM25+R	0.2350†	0.6328	0.8044	0.5000†	0.4963	0.4483	0.4318†	0.1936†	0.5198†	0.7447	0.4206†	0.4246†	0.3864†	0.3734†
	ANCE+VPRF-R	0.4300†	0.8177	0.8179	0.7037	0.7023	0.6790	0.6202†	0.4220†	0.8377	0.7958	0.7540	0.7472	0.6841	0.5760
	TCTV2+VPRF-R	0.4949†	0.8682	0.8942	0.7917	0.7464	0.7406	0.6876	0.4904†	0.8321	0.8655†	0.7817	0.7592	0.7144	0.6274†
	DistilBERT+VPRF-R	0.5156†	0.8606	0.8928	0.7731	0.7411	0.7387	0.6897	0.4974†	0.8899	0.9101†	0.8135	0.7973	0.7513	0.6535
	ANCE-PRF	0.4423†	0.8721	0.8293	0.7361	0.7204	0.7074	0.6270†	0.4340†	0.8881†	0.8286	0.8571†	0.7792	0.7275	0.5897
	TCTV2-PRF	0.4901†	0.8615	0.8888	0.7500	0.7606	0.7456	0.6802	0.4864†	0.8774	0.8562†	0.8254†	0.8038	0.7331	0.6252†
	DistilBERT-PRF	0.4996	0.8588	0.8968	0.7546	0.7648	0.7386	0.6778	0.4860	0.8810	0.8777	0.7976	0.7803	0.7306	0.6293
STRONG SIGNAL	BM25+R	0.3706†	0.8334†	0.8412	0.6505	0.6536†	0.6076†	0.5578†	0.3936†	0.8859†	0.8422	0.7695†	0.7040†	0.6411†	0.5566
	ANCE+VPRF-R	0.5119†	0.8816†	0.8531†	0.7708	0.7690	0.7511	0.6749†	0.5259†	0.9164†	0.8377	0.8132	0.8158	0.7670†	0.6398†
	TCTV2+VPRF-R	0.5769†	0.9136	0.9292	0.8002	0.7957	0.7773	0.7366†	0.6018†	0.9484†	0.9142†	0.8323	0.8227	0.7925†	0.6918†
	DistilBERT+VPRF-R	0.5931†	0.9410	0.9336	0.7967	0.8068	0.7971	0.7433	0.6022†	0.9738†	0.9255	0.8462	0.8260	0.7955†	0.7030
	ANCE-PRF	0.4907†	0.9060	0.8388	0.7986	0.7722	0.7484	0.6583†	0.4798†	0.8771	0.8241	0.7798	0.7685	0.7249	0.6078
	TCTV2-PRF	0.5326†	0.8807	0.9134	0.7654	0.7641	0.7600	0.7102	0.5348†	0.9220	0.8780†	0.8307	0.8223	0.7683†	0.6512†
	DistilBERT-PRF	0.5408†	0.8954	0.9124	0.7963	0.7821	0.7630	0.7041	0.5264†	0.9082	0.8915	0.8185	0.8084	0.7542	0.6561
MODERATE SIGNAL	BM25+R	0.2641	0.5738†	0.7979	0.4842†	0.4899	0.4871	0.4975	0.1960†	0.3855†	0.7956	0.3740†	0.3821†	0.4039†	0.4358†
	ANCE+VPRF-R	0.4365	0.8278	0.8477	0.7161	0.6951	0.6845	0.6444	0.3541†	0.6839†	0.8174	0.6177†	0.6286†	0.6091	0.5738
	TCTV2+VPRF-R	0.4675	0.7999	0.9088	0.7164†	0.7060	0.7052	0.6861	0.3754	0.5763†	0.8830†	0.5321†	0.5665†	0.5792†	0.6124
	DistilBERT+VPRF-R	0.4847	0.8143	0.9193	0.7029	0.7049	0.7075	0.6981	0.4075†	0.6412†	0.8889	0.5747†	0.6140†	0.6440	0.6315
	ANCE-PRF	0.4369†	0.7740	0.8324	0.6620	0.6905	0.6936	0.6385	0.3703	0.6882†	0.8143	0.6359†	0.6409†	0.6302	0.5627
	TCTV2-PRF	0.4841	0.8550	0.8977	0.7558	0.7403	0.7338	0.6915	0.4684†	0.8540	0.8583	0.7791	0.7545	0.7113	0.6355†
	DistilBERT-PRF	0.5049	0.8779	0.9069	0.7755	0.7472	0.7369	0.6900	0.4701	0.8215	0.8816	0.7229	0.7499	0.7018	0.6340
WEAK SIGNAL	BM25+R	0.1957†	0.3917†	0.7641	0.2118†	0.2518†	0.2902†	0.3643†	0.1839†	0.3712†	0.7433	0.2153†	0.2389†	0.2866†	0.3506†
	ANCE+VPRF-R	0.3915	0.7886	0.8130	0.6713	0.6526	0.6480	0.5836	0.3180†	0.6949†	0.7832	0.5575†	0.5609†	0.5250†	0.4968†
	TCTV2+VPRF-R	0.4408	0.8036	0.8875	0.6921†	0.6608	0.6692	0.6351	0.3440†	0.5926†	0.8352	0.4484†	0.4785†	0.4792†	0.5031
	DistilBERT+VPRF-R	0.4441	0.8022	0.8970	0.6667†	0.6622†	0.6561	0.6371	0.3904†	0.7441†	0.8603	0.5903†	0.5873†	0.5702†	0.5555†
	ANCE-PRF	0.3929	0.7121†	0.8159	0.5232†	0.5733†	0.5906†	0.5791	0.3594	0.7185†	0.7911	0.6181†	0.6142†	0.5935	0.5280
	TCTV2-PRF	0.4705	0.8487	0.8890	0.7315	0.7110	0.7053	0.6582	0.4703†	0.8827	0.8551†	0.7748	0.7488	0.7001	0.6079
	DistilBERT-PRF	0.5047	0.8757	0.9002	0.7639	0.7386	0.7262	0.6789	0.4746	0.8626	0.8747	0.7474	0.7675	0.6991	0.6213

Table 3: The Rocchio parameter settings for both datasets, with different PRF depths and different models.

	Depth		TREC DL 2019	TREC DL 2020
BOW	all	α	0.75	0.75
		β	0.15	0.15
Dense Retrievers	1	α	0.6	0.5
		β	0.4	0.5
	3	α	0.6	0.6
		β	0.4	0.4

change and show substantial improvements over all deep metrics, TCTV2+VPRF-Rocchio and DistilBERT+VPRF-Rocchio, on the other hand, also drops quickly. For using weak signals, BM25+Rocchio suffers more than 60% loss on several metrics compares to BM25, marginal improvements can be observed only for ANCE+VPRF-Rocchio and TCTV2+VPRF-Rocchio, where DistilBERT+VPRF-Rocchio still suffers from substantial loss. For the learnt PRF approaches, all three models show a more stable resilient of signal quality change, even with weak signal, the worst performance are just about 20% lower than the baseline.

3.2 Signal Quality and Representations

Next, we investigate how representations from different models impact effectiveness. For this analysis, we only consider the Rocchio PRF method (called Vector-PRF or VPRF-Rocchio for dense retrievers), as this is the only PRF method for which we have both bag-of-words and dense representations. We again refer to the results in Table 2.

For the bag-of-words representation (BM25 + Rocchio), effectiveness drops very quickly when moving from a strong signal to a weak signal: losses at times reach 80% for some metrics. This trend is observed across all datasets and metrics.

We now consider dense representations. ANCE+VPRF-Rocchio exhibits more stable behaviour with respect to changes of feedback signal quality than when bag-of-words are used: losses in the worst conditions are up to only 30%. However, TCTV2+VPRF-Rocchio and DistilBERT+VPRF-Rocchio show instead quite unstable patterns when changing the PRF signal quality: the methods suffer losses of more than 50% on some metrics when changing the feedback signal from strong to weak.

Our results suggest that better underlying representations, i.e. dense representations in place of bag-of-words representations, lead the same PRF technique to higher effectiveness, and this is regardless of the feedback signal quality. In fact, even with feedback signal of weak quality, losses obtained by the PRF mechanism on dense representations are lower than those obtained on bag-of-words representations. Differences do still exist however across the different dense representations, at least in the extent of th relative gains and losses depending on the quality of the PRF signal.

4 CONCLUSION

In this paper we conducted a systematic investigation of how the feedback signal quality impacts the effectiveness of pseudo relevance feedback for passage retrieval. We demonstrated that the strength of the PRF signals has a high impact on effectiveness; strong signals achieve higher gains in effectiveness, while weak signals hurt the effectiveness. However, we showed that the stability in performance differs from one PRF method to another. For instance, *learnt* PRF methods are more resilient to weak signals (noise) than not-learnt methods (e.g. Rocchio on either bag-of-words representations or dense retrievers – called VectorPRF Rocchio). We also showed that, under the same PRF method, dense representations are better than bag-of-words representations across all spectrum of feedback signal quality.

Our investigation is not without limitations. Importantly, we did not consider mixed signals (i.e., where the relevance of passages in the top k varies) and a broader set of PRF depths k. Mixed signals were not considered in this initial work so as to have a clear control of the signals and facilitate our investigation and results interpretations. However, in future work we plan to extend our analysis to more complex signals, including larger samples. In terms of feedback depth, we only studied $k = 1$ and $k = 3$ – although these are often popular settings, especially for the passage retrieval task, certainly they are not the only possible[5]. In addition, more PRF methods could have been investigated, including other neural PRF methods, e.g., ColBERT-PRF [15].

ACKNOWLEDGEMENTS

This research is funded by the Grain Research and Development Corporation (GRDC), project AgAsk (UOQ2003-009RTX).

REFERENCES

[1] Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query Expansion Techniques for Information Retrieval: A Survey. In *Information Processing & Management.*

[2] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 Deep Learning Track. In *Text REtrieval Conference, TREC.*

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2021. Overview of the TREC 2020 Deep Learning Track. In *Text REtrieval Conference, TREC.*

[4] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proceedings of The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[5] Victor Lavrenko and W Bruce Croft. 2017. Relevance-Based Language Models. In *ACM SIGIR Forum.*

[6] Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2022. Pseudo Relevance Feedback With Deep Language Models and Dense Retrievers: Successes and Pitfalls. In *Transactions of Information Systems.*

[7] Hang Li, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback: A Reproducibility Study. In *Proceedings of the 44rd European Conference on Information Retrieval.*

[8] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations. In *Proceedings of The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[9] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021).*

[10] Yuanhua Lv and ChengXiang Zhai. 2009. A Comparative Study of Methods For Estimating Query Language Models with Pseudo Feedback. In *Proceedings of the 18th ACM ACM International Conference on Information and Knowledge Management.*

[11] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. In *Workshop on Cognitive Computing at NIPS.*

[12] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. In *Foundations and Trends in Information Retrieval.*

[13] J.J. Rocchio. 1971. Relevance Feedback in Information Retrieval. In *The SMART Retrieval System - Experiments in Automatic Document Processing.*

[14] Chih-Fong Tsai, Ya-Han Hu, and Zong-Yao Chen. 2015. Factors Affecting Rocchio-based Pseudo-Relevance Feedback in Image Retrieval. In *Journal of the Association for Information Science and Technology.*

[15] Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2021. Pseudo-Relevance Feedback for Multiple Representation Dense Retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval.*

[16] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning For Dense Text Retrieval. In *arXiv preprint arXiv:2007.00808.*

[17] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. In *Journal of Data and Information Quality (JDIQ).*

[18] HongChien Yu, Chenyan Xiong, and Jamie Callan. 2021. Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management.*

[5] While some dense retrieval based PRF methods are limited in terms of the maximum number of passage they can consider as feedback [7, 18], others are not [6].