The Clearest Pseudo Relevance Feedback

Bryant Curto
University of Waterloo
Canada

ABSTRACT

We propose and demonstrate a novel language modeling approach to pseudo relevance feedback that deviates from methods previously proposed. To do this, we make use of clarity score as proposed by Townsend et al. In this endeavor, we reproduce some of the results of Townsend et al. in their paper proposing clarity score. While we were not able to reproduce their results, and therefore were not able to evaluate our method, we identify an issue that we believe to be the cause and propose future work.

ACM Reference Format:
Bryant Curto. 2021. The Clearest Pseudo Relevance Feedback. In *Proceedings of Winter '21 CS848 High Recall IR.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

The language modeling approach to information retrieval was first introduced by Ponte and Croft and has been shown to perform well [**?**]. In this approach, each document is viewed as a language sample and a query as a generation process. With this conception, documents can be ranked based on the probabilities of producing a given query from the corresponding document language models.

Relevance feedback is the method of getting feedback from the user on one or more initial sets of results in order to improve the final set of results. Relevance feedback operates on the belief that a user may have difficulty formulating a good query when they do not know the collection well, but is able to determine whether or not a specific document is relevant. By asking the user about the relevance (or non-relevance) of an intermediate set of results, the system can use this feedback to improve the final results.

Pseudo relevance feedback, also known as blind relevance feedback, automates the manual part of relevance feedback so the user receives the final set of results without additional work. In general, pseudo relevance feedback methods retrieve an initial set of documents, assume that the top ranked documents are relevant, and then use this relevance feedback to retrieve another set of documents. This process can continue for a number of iterations before eventually returning a final set of results back to the user.

In this work, we propose a novel language modeling approach to pseudo relevance feedback that deviates from those methods previously proposed by making use of clarity score as proposed by Townsend et al. [**?**]. Several methods for performing pseudo relevance feedback are based on a language modeling approach by augmenting the language model generated from the query with a model derived from the feedback documents. In our proposed method, a clarity score is computed from the feedback documents of a query, and the top contributing terms to the clarity score are used to augment the query – similar to query expansion. This augmented query is then used to retrieve the result set. We believe that our method is simpler and can be applied more generally.

Winter '21 CS848 High Recall IR, April 30, 2021, Waterloo, Canada
2021.

2 BACKGROUND

2.1 Clarity Score

Clarity score is a post-retrieval prediction method. It is a measure of the degree of ambiguity of a query with respect to a collection of documents. More specifically, it is a measure of the degree of dissimilarity between the language usage associated with the query and the generic language of the collection as a whole. It is closely related to the lack of ambiguity and thus is called the clarity score.

The Clarity score is simply the KL-divergence between the query language model (i.e., the language model of the result set) and the collection language model (i.e., the language model of the entire collection). Each of these language models is a unigram language model, which is a probability distribution over the terms in each of their respective document collections.

Let q denote a query, d denote a document, D denote the entire collection of documents, and D_q denote the set of documents retrieved by query q. With this notation, let $P(\cdot|D_q)$ denote the query language model and $P(\cdot|D)$ denote the collection language model. Then, the Clarity score is defined as:

$$\begin{aligned} Clarity(q) &= KL_{div}(P(\cdot|D_q) \,\|\, P(\cdot|D)) \\ &= \sum_{t \in V(D)} P(t|D_q) \, \log \frac{P(t|D_q)}{P(t|D)} \end{aligned} \tag{1}$$

where $V(D)$ is the vocabulary of the entire collection and t is a term in this vocabulary.

The collection language model is assumed to be computed directly from the collection (i.e., without estimating) because it is assumed that the collection will not change. Therefore, the collection language model only needs to be computed once.

In contrast, the query language model is estimated using Method 1 proposed by Lavrenko and Croft [**?**]. It is computed as follows:

$$P(t|D_q) = \sum_{d \in D_q} P(t|d) \, P(d|q) \tag{2}$$

$P(t|d)$, the probability of term t in some document d, is estimated by the relative frequency of t in d, smoothed by a linear combination with its relative frequency in the collection:

$$P(t|d) = \lambda \frac{tf(t,d)}{|d|} + (1-\lambda) \, P(t|D) \tag{3}$$

Townsend et al. set λ to 0.6. For all of our experiments, we do the same.

$P(d|q)$, the probability of a document given the query, is obtained through Bayesian inversion with uniform prior probabilities for documents in D_q and zero for prior probabilities for documents not in D_q:

$$P(d|q) = \frac{P(q|d) \, P(d)}{\sum_{d' \in D_q} P(q|d') \, P(d')} \tag{4}$$

Finally $P(q|d)$, the probability of a query given a document, is computed using the query-likelihood unigram language modeling

approach [**?**]:

$$P(q|d) = \prod_{t \in q} P(t|d) \tag{5}$$

In Townsend et al.'s paper proposing the clarity score, a set of up to 500 unique documents, all containing at least one query term, were used in place of the documents retrieved by a query, D_q. In a later work, the same authors find that the top 500 retrieved documents could be used following the observation that $P(d|q)$ generally decreases sharply before this cutoff [**?**]. For ease, we use the latter of these methods in our experiments.

3 RELATED WORK

There has been much work in using language models to perform pseudo relevance feedback, but none that we can find using language modeling or the clarity score as we have. The two representative approaches for using language modeling to perform pseudo relevance feedback are the relevance model [**?**] and the mixture model [**?**]. For both of these methods, feedback documents are used to estimate a better query language model.

A relevance model [**?**] makes the assumption that each query term is generated from the relevance model $P(w|R)$, where R is the class of documents that are relevant to a user's information need. However, it is impossible to know R without training data. Thus, the relevance model is estimated by assuming that the top-ranked feedback documents are samples from the relevance model. More specifically, the probability of a term in the relevance model is estimated by its probability in each feedback documents weighted by the correspondence of the feedback document to the query. The estimated relevance model is then combined with the original query model to form an updated query language model.

In mixture-model feedback [**?**], the words in the set of feedback documents are assumed to be drawn from two models: a background language model and a query topic language model. Mixture-model feedback is used to estimate the query topic model that most closely matches that of the feedback documents. It does this by first separating the query topic model from the background model, which can be thought of as background "noise". The query topic model is then combined with the original query model to form an updated query language model.

For both of these methods, documents are then scored by computing the KL-divergence between the updated query model and the language model of each document. The top scoring documents are then returned.

These methods require the system designer to decide a priori the degree to which the updated query model is influenced by the estimated relevance model and the query topic model, respectively. Further, these methods raise a few questions. Would it be better to expand the query with a few topical terms? What can we do if we want to use a non-language-modeling based approach for the re-retrieval? We believe that our proposed method can provide answers to these questions.

4 CLARITY SCORE AND AVERAGE PRECISION

As a starting point, we reproduce some of the evaluations performed by Townsend et al. as a means of validating our implementation of the clarity score equation. To do this, we compute correlation between clarity scores and average precision scores for several TREC Ad Hoc Track test collections and queries.

Collection	Topics	Source	Num Queries	Corr	P-value
TREC-5	251-300	title	50	0.459	6.5×10^{-4}
TREC-7	351-400	title	50	0.577	2.7×10^{-5}
TREC-8	401-450	title	50	0.494	2.7×10^{-4}

Table 1: Some of the evaluation results of Townsend et al. in [?] showing the correlation of clarity score with average precision in several TREC Ad Hoc test collections using topic titles as queries.

To compute clarity scores, language models for each collection and for each document in each collection needed to be computed. To do this, the handyman tool that comes with wumpus was used [**?**]. Wumpus is an information retrieval system developed at the University of Waterloo by Stefan Büttcher. While its main purpose is to study issues that arise in the context of indexing dynamic text collections in multi-user environments, we use it in our work for performing information retrieval on static text collections. Further, all document retrievals are performed using wumpus.

As in the original paper, we use the Spearmann rank correlation test to compute the correlation. To perform this test, two rankings of queries are created where one is ordered according to the queries' clarity score and the other by their average precision. A correlation of 1 indicates perfect agreement between the rankings and a correlation of -1 indicates perfect disagreement (i.e., the rankings are inverses of each other). The p-value indicates the estimated probability that an apparent correlation as extreme, or more extreme, would occur by chance if clarity scores and average precision scores were not associated.

Some of the evaluation results presented in the original paper can be seen in Table 1. Reviewing the columns from left to right, *Collection* refers to the TREC Ad Hoc Track whose document collection was used for the retrieval. *Topics* refers to the range of topic numbers from which queries were generated. *Source* refers to the component of the topic (e.g., title, description, narrative) from which the query was derived. *Num Queries* refers to the number of queries from which evaluation results are derived. *Correlation* and *P-value* refer to the evaluation measures previously described. As noted in the original paper, the results show a strong positive association between the clarity score and the average precision of a query.

Our correlation results can be seen in Table 2 following a similar table layout. The first column, *Retrieval Method*, refers to the wumpus ranked retrieval method used for the document retrieval. BM25 refers to Okapi BM25. LM refers to a language modeling based approach for ranked retrieval. The specific method is based on Bayesian smoothing with Dirichlet priors as proposed by Zhai and Lafferty [**?**]. Note also that we source our queries from not only topic titles but also topic descriptions. Further, note that we compute rank correlations from queries derived from topic titles only, topic descriptions only, and topic titles or descriptions. (Across

Retrieval Method	Collection	Topics	Source	Num Queries	Corr	P-value
BM25	TREC-5	251-300	title	50	-0.01878	0.84744
			description	50	-0.01388	0.88692
			title or description	100	0.04606	0.49713
	TREC-6	301-350	title	50	**0.34857**	$\mathbf{3.6 \times 10^{-4}}$
			description	50	-0.04327	0.65752
			title or description	100	0.06465	0.34059
	TREC-7	351-400	title	50	-0.02857	0.76970
			description	50	0.02531	0.79540
			title or description	100	-0.02263	0.73872
	TREC-8	401-450	title	50	-0.02367	0.80833
			description	50	-0.03020	0.75694
			title or description	100	-0.09535	0.15982
LM	TREC-5	251-300	title	50	-0.13470	0.16753
			description	50	0.01878	0.84744
			title or description	100	0.02061	0.76131
	TREC-6	301-350	title	50	-0.06122	0.53042
			description	50	**0.18857**	$\mathbf{5.333 \times 10^{-2}}$
			title or description	100	-0.01212	0.85818
	TREC-7	351-400	title	50	-0.07592	0.43661
			description	50	0.07429	0.44654
			title or description	100	-0.05616	0.40772
	TREC-8	401-450	title	50	0.08245	0.39819
			description	50	0.01061	0.91341
			title or description	100	0.01657	0.80707

Table 2: The correlation of clarity score with average precision (and p-values) for several TREC collections using several retrieval methods.

all topics used, titles consist of 3.04 terms on average. Meanwhile, descriptions consist of 15.865 terms on average.)

Inspecting our results (Table 2), there appears to be no association between the clarity score and average precision of a query in contrast to what was observed in the original paper (Table 1). Looking at the p-values, our results indicate a rather high probability (i.e., over 50% in most cases) that an apparent correlation as extreme, or more extreme, would occur by chance if clarity scores and average precision scores were not associated. The only exceptions are the TREC-6 title derived queries when BM25 was used and the TREC-6 description derived queries when LM was used (in bold for convenience). However, it is important to note that neither of these correlations are as large or these p-values as small as those presented in the original paper.

To better understand our results, we compute the rank correlation between clarity score and average precision per topic using the queries generated from each topic's title and description. This is to see if correlation results observed in the original paper can be replicated by comparing queries about the same topic. The results of this evaluation can be seen in Table 3. *Num Queries* indicates the number of queries with which each rank correlation was computed. *Correlation* shows the mean of the rank correlations. For each retrieval method and collection, 50 rank correlations, each of which was computed using 2 queries generated from the same topic, were averaged. Rank correlations are either 1.0 or -1.0, so we do not show standard deviation. These results indicate that, even when looking at the correlation between clarity score and average precision at the per-topic level, the association is tenuous at best.

Inspecting the clarity scores directly, we come to an interesting discovery. Averages and standard deviations of clarity scores and their top term contributions are presented in Table 4. The top term contribution is the term $t \in V(D)$ that contributes the most to the clarity score of a query as seen in Equation 1. Clarity scores appear to cluster around 0.8 for all collections. Further, top terms contributions account for approximately 10% of the clarity score.

We believe that this is notable because we implemented the clarity score equation in Python using its built-in `float` type. We believe that the lack of association between clarity score and average precision observable in our results may be attributed to our using the inexact `float` type to compute clarity scores. In conjunction with the smallness of each term's contribution, we believe that the error introduced in our calculations may have resulted in computed clarity scores deviating far enough from their actual values to have broken the association between clarity score and average precision.

5 CLARITY BASED RELEVANCE FEEDBACK

In this section, we demonstrate how one might perform our proposed method of using clarity score for pseudo relevance feedback.

Using the queries from the previous section that showed the strongest association between clarity score and average precision – TREC-6 titles retrieved using BM25 – we now attempt to use clarity score for pseudo relevance feedback. To do this, we reformulate

Retrieval Method	Collection	Topics	Num Queries	Corr
BM25	TREC-5	251-300	2	0.04
	TREC-6	301-350	2	-0.2
	TREC-7	351-400	2	0.0
	TREC-8	401-450	2	0.12
LM	TREC-5	251-300	2	0.0
	TREC-6	301-350	2	-0.16
	TREC-7	351-400	2	-0.08
	TREC-8	401-450	2	0.08

Table 3: The correlation between clarity score and average precision using the queries generate from a topic's title and description averaged over topics for several TREC collections using several retrieval methods.

Retrieval Method	Collection	Topics	Source	Clarity Score	Top Contributing Term
BM25	TREC-5	251-300	title	0.86634 ± 0.37899	0.08686 ± 0.05614
			description	0.79630 ± 0.38183	0.07592 ± 0.06160
	TREC-6	301-350	title	0.79440 ± 0.35243	0.08990 ± 0.06049
			description	0.89668 ± 0.32684	0.08227 ± 0.05028
	TREC-7	351-400	title	0.77408 ± 0.33313	0.09173 ± 0.04681
			description	0.92380 ± 0.36200	0.08527 ± 0.05530
	TREC-8	401-450	title	0.79003 ± 0.33647	0.10026 ± 0.06546
			description	0.91644 ± 0.34736	0.09584 ± 0.05961
LM	TREC-5	251-300	title	0.85377 ± 0.37518	0.08592 ± 0.05516
			description	0.85377 ± 0.37518	0.07530 ± 0.06089
	TREC-6	301-350	title	0.79554 ± 0.35534	0.09042 ± 0.06086
			description	0.90732 ± 0.32074	0.08388 ± 0.05123
	TREC-7	351-400	title	0.74949 ± 0.34559	0.08952 ± 0.04795
			description	0.92628 ± 0.35919	0.08571 ± 0.05474
	TREC-8	401-450	title	0.78389 ± 0.33991	0.09983 ± 0.06596
			description	0.91574 ± 0.34561	0.09582 ± 0.05962

Table 4: The mean and standard deviation of clarity scores and top term contribution to each clarity score for several TREC collections using several retrieval methods.

each query by adding the term the contributed the most to a given query's clarity score.

For example, topic 302, which received a clarity score of 1.40265, has the title, *"Poliomyelitis and Post-Polio"*, and the following description: *"Is the disease of Poliomyelitis (polio) under control in the world?*

The term that contributed most to the clarity score of the query generated from this title was *"health"* with a term contribution of 0.10974. Therefore, the newly formulated query includes the term *"health"*.

Table 5 shows the mean and standard deviation of the increase in average precision from using queries augmented with clarity based pseudo relevance feedback as compared to the original queries. As may have been expected, average precision remains approximately the same or even goes down a bit.

Retrieval Method	Collection	Topics	Source	Num Queries	Average Precision
BM25	TREC-6	301-350	title	50	-0.00892 ± 0.09378

Table 5: Mean and standard deviation of increase in average precision from augmenting queries with the term contributing most to the query's clarity score.

6 FUTURE WORK

As discussed in Section 4, we believe that the observed lack of association between clarity score and average precision may be a result of our use of Python's built-in inexact `float` type. As future work, we plan on re-implementing our tools to compute clarity scores using a type that can represent numbers exactly throughout the computation (i.e., arbitrary precision rational number types).

With this correction in place, we plan on reevaluating the correlation between clarity score and average precision. We also intend on reevaluating our pseudo relevance feedback method on a wider range of queries and by augmenting queries with a varying number of top contributing terms.

While pseudo relevance feedback techniques generally improve retrieval performance on average, they are not robust. They tend to advantage some queries and disadvantage other queries [**? ?**]. This limits the usefulness of pseudo relevance feedback in real world retrieval applications. As future work, we are interested in

understanding to what degree this issue impacts our proposed method.

7 CONCLUSION

In this work, we propose and demonstrate a clarity score based pseudo relevance feedback method that we believe is simpler and can be applied more generally than other existing language modeling based approaches. To do this, we attempted to reproduce the results of Townsend et al. showing the correlation between clarity score and average precision in their paper proposing clarity score [?]. While we were not able to reproduce their evaluation results, we have identified an issue in our setup that we believe, once fix, will enable us to get the expected result and will enable us to fully evaluate our proposed method.

8 ACCESSIBILITY

All code can be found (here).

REFERENCES

[] Stefan Buettcher. 2009. The Wumpus Information Retrieval System—On-disk index data structures.

[] Kevyn Collins-Thompson. 2009. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 837–846.

[] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting Query Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland) *(SIGIR '02)*. Association for Computing Machinery, New York, NY, USA, 299–306. https://doi.org/10.1145/564376.564429

[] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2006. Precision Prediction Based on Ranked List Coherence. *Inf. Retr.* 9, 6 (Dec. 2006), 723–755. https://doi.org/10.1007/s10791-006-9006-4

[] Donna Harman and Chris Buckley. 2004. The NRRC reliable information access (RIA) workshop. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 528–529.

[] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 120–127. https://doi.org/10.1145/383952.383972

[] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 275–281. https://doi.org/10.1145/290941.291008

[] Fei Song and W. Bruce Croft. 1999. A General Language Model for Information Retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management* (Kansas City, Missouri, USA) *(CIKM '99)*. Association for Computing Machinery, New York, NY, USA, 316–321. https://doi.org/10.1145/319950.320022

[] Chengxiang Zhai and John Lafferty. 2001. Model-Based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (Atlanta, Georgia, USA) *(CIKM '01)*. Association for Computing Machinery, New York, NY, USA, 403–410. https://doi.org/10.1145/502585.502654

[] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179–214. https://doi.org/10.1145/984321.984322