Impact of Pseudo Relevance Feedback on Personalized Headline Generation

Sanath Upadhya
University of Massachusetts, Amherst
Amherst, MA, USA

ABSTRACT

Large language models (LLMs) are being used in a variety of novel ways currently, be it in summarizing a document or generating a news headline from an article. The personalization of the large language models based on the user's preferences/history would theoretically help in generating better results. This paper looks at the effectiveness of personalizing the large language models for automatic news headline generation, by prompting the large language model with personalized input that uses pseudo relevance feedback. The paper also compares the different algorithms of pseudo relevance feedback on the final output by the LLM.

KEYWORDS

Information retrieval, pseudo relevance feedback, large language models, LaMP benchmark

ACM Reference Format:
Sanath Upadhya. 2023. Impact of Pseudo Relevance Feedback on Personalized Headline Generation. In *Proceedings of CS646 Fall 2023 Final Project (CS646 Final Project).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

The advent of large language models (LLM) is promising to revolutionize various aspects of human life, be it looking for a specific information or generating headlines for a given article. It is assumed that, with better/relevant input prompts, i.e. the query string that the user gives to the large language model (LLM), the output generated by the LLM would also improve.

In order to verify if this is indeed the case, LaMP benchmark offers different datasets,where, each data item within it has the following structure [1]:

Input data, expected output data, input profile

The input profile contains all data that is associated with a particular user. It should be noted that, due to the context length constraint seen in large language models, it is infeasible to give the entire input profile as the query to the LLM. Thus, there is a need to only select a subset of these input profiles and only feed the relevant ones to the LLM as part of the input query.

In this paper, we analyse the impact of using different pseudo relevance feedback algorithms/settings on the final output that is generated by the LLM. We focus our tests only on the type of queries that ask the LLM to generate a headline for the given article. The dataset that is taken from LaMP benchmark (LaMP 4: Personalized News Headline Generation) is of the following type:

Article, Headline,(set of article:headline written by same user)

A retrieval model, that takes the article for which the headline has to be generated is run on the input profile, the top k documents from the input profile is selected and fed into the large language model, along with the input article and the performance of the LLM is measured.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CS646 Final Project, Dec 12, 2023, Amherst, MA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2 RELATED WORK

There has been work done on summarizing the documents based on pseudo relevance feedback and user preferences [2][3]. However, the techniques mentioned in these papers are used for summarizing the documents. The length of the articles that are considered in these papers is much larger than the ones present in LaMP benchmark, and hence, there connot be a direct one-to-one comparison between the results obtained in this paper and the ones mentioned in the paper above.

3 PROBLEM STATEMENT

This paper mainly analyses the impact of having pseudo relevance feedback in the retrieval model, while selecting the most appropriate data item/s from the input profile.

This input profile is then given as the input query (in addition to the article whose headline has to be generated) to the LLM and the output generated by the LLM is measured with the expected output, which is present in the LaMP benchmark.

The metric that is used to measure the similarity between the two strings (the actual headline generated by the user and the headline generated by the LLM) using Rouge-1 and Rouge-L scores [5].

4 METHODOLOGY

The code for the below experiments can be found at:
https://github.com/sanath-upadhya/prf-lamp.

The paper tests two flavors of relevance models - RM1 and RM3 for retrieving the most appropriate results from the input profile, based on the input article. For each of the above algorithm, we modify the number of retrieved documents that is considered while doing pseudo relevance feedback (PRF), the number of terms that are used in PRF and the weight assigned to the original term.

Thus, we have the following algorithms that are tested [6]:

(1) RM1
(2) RM3

The number of documents that are considered during PRF, the number of terms that are used during PRF and the weight for the original query that are used during the experimentation are as follows:

(1) Documents = 3, Terms = 10, Weight = 0.25
(2) Documents = 3, Terms = 10, Weight = 0.75
(3) Documents = 2, Terms = 5, Weight = 0.25
(4) Documents = 1, Terms = 5, Weight = 0.25

The number of terms that are to be used in PRF is varied between five and ten. This is because, the average input length (i.e., the article length) is around thirty, and since we try to find the most similar article-headline pair from the input profile, we use the number of terms in PRF to be substantially smaller than the input length. The number of documents to be considered for PRF is varied between one, two and three, as it is mentioned that there is a minimum of four articles that each user has written.

We also see what would be the impact on the final output generated by the LLM, if the number of user profiles that were fed into the LLM is increased.

For each set of experiments, we randomly select one hundred (100) input articles from the data set, retrieve it's input profile and get the most similar <article-headline> pair/s by running a retrieval model that uses one of the PRF relevance models within it.

The input query that has to be fed into the LLM is constructed as follows:

Generate a headline for the following article: <input article>

If the experiment has any input profile that has to be appended to the above, the input profile is constructed as follows:

The user has previously generated the headline <input-profile-headline> for the article <input-profile-article>

The output generated by the LLM (we use the FLAN-T5-large LLM) [4] is then compared with the ground truth string, i.e. the headline which was generated by the user. The metric that is used to compare the two strings are Rouge-1 and Rouge-L, which are the same metric used in the LaMP benchmark [1].

We also run the same experiment and retrieve results from the input profile without using pseudo relevance feedback, get the output from the LLM and compare this output with the ground truth output.

5 RESULTS

The results of the experiments are as mentioned in the table below:

As can be clearly seen from Table 1, the performance of the large language model (flan-t5-large) decreases when we provide additional inputs from the input profile (i.e., the performance of this LLM decreases with personalization). We also see that the number of non-zero results (which indicates that atleast some portion of the output generated by the LLM is valid) also decreases with an increase in the number of retrieved result. This is because, as flan-t5-large has a modest number of nodes (compared to the LLM of GPT4 and flan-t5-xxl), and as a result of this, it is not capable of keeping the entire context that was presented in the input string.

However, we can still compare the impact of pseudo relevance feedback on the output generated by LLM. The output generated by different algorithms of PRF when the total number of articles that is fed into the LLM is two, is as shown in Table 2. As can be clearly seen from the table, the algorithm RM3 performs better than RM1, and we also see that, the optimal performance of RM3 is when the total number of feedback terms is small and the total number of documents that are considered to generate these terms are also small. We also see that, in such a scenario, the performance of the LLM is better than that of the LLM without any pseudo relevance feedback.

Table 3 indicates the output of the LLM when the total number of input profile that is fed into it is one. We see that, RM3 again performs better than RM1, and the optimal performance of RM3 is when the number of terms (used in relevance feedback) is small and the total number of documents that are to be considered to generate these terms are also small. We again see that, the performance of the LLM is slightly better in this situation when compared to the base case (of not having pseudo relevance feedback in the retrieval model).

We also see that a substantial amount of the output generated by the LLM has the Rouge-1 and Rouge-L value (when compared with the ground truth) to be zero. This number decreases as the number of input profiles fed into the LLM decreases. This can be a result of the LLM getting confused by the large input query, i.e. the LLM is unable to keep the context within it.

6 CONCLUSION

As can be clearly seen from the results, incorporating the pseudo relevance feedback in out retrieval model for getting the most appropriate input profile increases the overall performance of the large language model (LLM) under test. However, adding more input profiles to the input query of the large language model degrades the performance of the model for the headline generation task. Thus, the importance of choosing the right relevance model when using pseudo relevance feedback becomes critical.

It is also noticed that, there is a substantial increase in the output metric (i.e., Rouge-1 and Rouge-L scores) when we discard all the outputs that have a zero value for these. This indicates that, a better output would be expected from the large language model (LLM) if we identify such inputs before feeding the input query to the LLM. One such technique might be to have a cut-off value for displaying the results of the input profile. If no relevant headline-article pair is found in the input profile, it would be better to not include any personalization data.

In the dataset (LaMP benchmark, dataset 4) that was used to conduct the experiments, we found that the best performance was achieved when the retrieval model was using pseudo relevance feedback, with the number of terms being used is five and the number of documents used to retrieve these terms to be one.

7 NEXT STEPS

This paper only focused on a single large language model and could measure the impact of pseudo relevance feedback on headline generation on this model. Some of the open ended questions that are still relevant are as follows:

(1) Is the impact of pseudo relevance feedback on news headline generation the same across the large language models?
(2) Is there an optimum number of terms that has to be used in PRF models so that the output generated by the LLM matches the user's style?
(3) The impact of other algorithms that are used to do pseudo relevance feedback on th eoutput of the LLM.

8 REFERENCES

(1) Salemi, Mysore, Bendersky and Zamani. "LaMP: When Large Language Models Meet Personalization." (2023).
(2) Park, Cha, and Kwon. "Personalized Document Summarization Using Pseudo Relevance Feedback and Semantic Feature." (2014)
(3) Wang, Pan, He, Huang, Wang, and Tu. "A Pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval." (2020)
(4) Chung, Huou, Longpre et al."Scaling Instruction-Finetuned Language Models." (2022)
(5) Yamada, Hitomi, Tamori, Sasano, Okazaki, Inui, Takeda. "Transformer based lexically constrained headline generation." (2021)
(6) Cao, Nie, Gao, Robertson."Selecting good expansion terms for pseudo-relevance feedback." (2008)

Received 12 December 2023; revised 12 December 2023; accepted 12 December 2023

Table 1: Results with no PRF

Retrieved result from profile	Non-zero results	Non-zero Rouge-1	Non-zero Rouge-L	Overall Rouge-1	Overall Rouge-L
0	79	0.2445	0.2173	0.1858	0.1651
1	68	0.2149	0.1883	0.1461	0.1280
2	60	0.2093	0.1761	0.1256	0.1056

Table 2: PRF, Retrieved Results = 2

Algorithm	Number of docs	Number of feedback terms	Original weight	Non-zero results	Non-zero Rouge-1	Non-zero Rouge-L	Overall Rouge-1	Overall Rouge-L
RM1	3	10	NA	56	0.1750	0.1587	0.0980	0.0888
RM1	2	5	NA	66	0.2251	0.2090	0.1485	0.1379
RM1	1	5	NA	46	0.1993	0.1868	0.0917	0.0859
RM3	3	10	0.25	67	0.1817	0.1769	0.1199	0.1167
RM3	2	5	0.25	60	0.1900	0.1789	0.1102	0.1038
RM3	1	5	0.25	57	0.2587	0.2535	0.1448	0.1419
RM3	3	10	0.75	56	0.1952	0.1702	0.1093	0.0953

Table 3: PRF, Retrieved Results = 1

Algorithm	Number of docs	Number of feedback terms	Original weight	Non-zero results	Non-zero Rouge-1	Non-zero Rouge-L	Overall Rouge-1	Overall Rouge-L
RM1	3	10	NA	73	0.2648	0.2434	0.1906	0.1752
RM1	2	5	NA	64	0.2215	0.2102	0.1373	0.1303
RM1	1	5	NA	56	0.2224	0.1967	0.1245	0.1101
RM3	3	10	0.25	57	0.2451	0.2380	0.1372	0.1332
RM3	2	5	0.25	69	0.1972	0.1810	0.1341	0.1231
RM3	1	5	0.25	73	0.2222	0.1889	0.1600	0.1360
RM3	3	10	0.75	62	0.1935	0.1879	0.1200	0.1165