Information Retrieval Combined with Large Language Model: Summarization Perspective

Shivani Choudhary, Niladri Chatterjee, Subir Kumar Saha
{shivani@sire, niladri@maths, saha@mech}.iitd.ac.in
Indian Institute of Technology Delhi
Hauz Khas, Delhi-110016, India

Abstract

Conventional information retrieval procedures typically entail multiple stages, encompassing information retrieval and subsequent response generation. The quality of the response derived from the retrieved content significantly influences the overall efficacy of the retrieval process. With the advent of large language models, it is possible to utilize larger contexts to generate more cogent summaries for users. To ensure the production of contextually grounded and pertinent responses, particularly in conversational models, a good retrieval mechanism acts as a keystone. This study aims to develop a conversational engine adept at extracting relevant documents and generating pertinent responses by summarizing key passages, leveraging various types of language models.

1 Introduction

Information retrieval, conventionally, involves retrieving a list of passages or documents from a large corpus based on their relevance to the user's query. This process can be executed as a single-stage or multi-stage pipeline. In a single-stage pipeline, documents or passages are retrieved from the corpus based on their relevance score, which can be calculated using metrics such as BM25 (Robertson and Zaragoza, 2009) score, vector similarity score, or a combination of different evaluation parameter, such as nDCG, P@K.

Recently, information retrieval has been adapted to conversational settings, where systems are tasked with responding to a sequence of user questions. The subsequent question in this series may not necessarily be a direct continuation of the preceding one. However, users retain the freedom to transition between different conversational contexts. Therefore, it becomes imperative to retain contextual information from previous interactions.

2 Problem Description

The iKAT-2023 aimed to assess user responses as conversation threads diverge into different paths following several initial turns, often referred to as ice breakers. Conversational turns within a thread are categorized into two types: manual and automatic. The automatic thread simulates human responses more realistically, incorporating ambiguity related to the use of context and pronouns. In contrast, the manual thread involves resolving co-references, such as pronouns, manually, with human intervention in reformulating queries.

In addition to queries, a *Personal Textual Knowledge Base* (PTKB) containing supplementary information, which may or may not be relevant for generating a response to the query, is provided. iKAT-2023 comprises two participation tracks. In the automatic track, participants are prohibited from utilizing any part of manually resolved queries and ground truth PTKB statements. Conversely, the manual track grants participants the flexibility to utilize ground truth PTKB statements.

A conversation thread S can be defined as a series of utterances $\{u_1, u_2, ..., u_n\}$ as turn. Each of turn u_i may or may not be related to the previous thread.

- Build a document retrieval engine that can return a list of relevant passages based on user query.
- Selection and ranking of PTKB relevant to the input query
- Investigate the quality of the summarization.

Dataset: ClueWeb-22-B [1]

[1] https://www.trecikat.com/

3 Motivation and Method

The information retrieval approach within conversational threads serves as a crucial focal point. Typically, the information retrieval pipeline adopts a multi-step structure (Choudhary, 2022). The initial stage of this pipeline involves retrieval, which can take the form of sparse retrieval (a variant of term frequency match), dense retrieval (based on continuous representations generated by neural models), or a hybrid approach. Subsequently, this initial stage is followed by a re-ranking stage, often facilitated by a language model based on transformer architecture, such as T5 or BART (Vaswani et al., 2017; Raffel et al., 2019; Lewis et al., 2019). The re-ranker model is trained with the objective of assigning a high relevance score to relevant query-passage pairs and a low score to negative samples.

We have established a multi-step retrieval pipeline, illustrated in Figure 1. This pipeline uses ClueWeb-22B document corpus.Vector embedding for ClueWeb-22B are generated for dense retrieval using, *all-MiniLM-L6-v2* as an encoder for our corpus. Passges were created using the official code for passage chunking available at [2]. FAISS(Johnson et al., 2019) acted as vecto store for embeddings. For retrieving passages, datasets from [3] were utilized. To enhance retrieval speed, we partitioned the main index into 8 sub-indexes, allowing for efficient multi-processing.

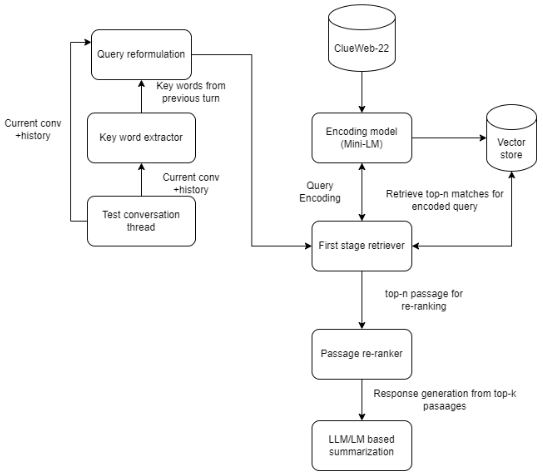

Figure 1: Retrieval pipeline for experiment. Desnse retrieval indexes are set up for Clueweb-22 B (Overwijk et al., 2022)

In a conversational setting, maintaining the coherence and central theme of the conversation across turns poses a challenge. Often, deviations during the conversation can obscure the primary points being discussed.

Pre-trained model	bart-large-cnn-samsum
Dataset	CANARD[4]
Learning rate	1e-5
FP16	True

Table 1: Details of the BART fine-tuning for summarization

To address this issue and preserve the contextual information of a chat session S, we employ a keyword extraction module. This module analyzes each utterance u_i along with the conversation history u_1 ... u_{i-1} to identify the most relevant top-k n-grams. This is achieved using a language model backbone, with BERT and FLAIR being the chosen models (Devlin et al., 2019; Akbik et al., 2019). For this purpose, we utilize KeyBERT [5].

Automatic queries necessitate query reformulation before retrieval can be initiated. To accomplish this, we employ BART and T5-based models to summarize the conversation up to the current turn. This summarization process integrates the conversation history, current query, relevant statements from the Personal Textual Knowledge Base (PTKB), and keywords to generate a reformulated query. This reformulated query serves as an approximate summary of the conversation up to the current turn.

The reformulated query is then utilized for retrieval. Using the *all-MiniLM-L6-v2* encoder, depicted in Figure 1, the reformulated query is encoded to retrieve passages from the vector store. Subsequently, a language model-based re-ranking engine, as proposed by (Zhang et al., 2022; Pradeep et al., 2021), analyzes the reformulated query and retrieved passages from the first stage to generate relevance scores for query-passage pairs.

To generate the final response for presentation to the user, various engines are employed to summarize the top-k paragraphs. This process involves conventional language models as well as the large language model LLAMA-7B introduced by (Touvron et al., 2023).

3.1 Fine-tunning procedure

To generate reformulated queries, we fine-tuned a BART-based language model pre-trained on the

[2] https://github.com/irlabamsterdam/iKAT

[3] https://huggingface.co/docs/datasets/index

[5] https://github.com/MaartenGr/KeyBERT

Run Name	P@5	nDCG@5	success@1	sucess@10
run_automatic_dense_monot5	0.3545	0.2361	0.3864	0.6591
run_automatic_dense_damo_canard_16000	0.3057	0.2122	0.3125	0.5625
run_automatic_llm_damo	0.2625	0.1737	0.3068	0.5795
run_automatic_dense_mini_LM_reranker	0.2409	0.1502	0.2500	0.5000

Table 2: Performance of the system compared against non-pruned qrels

Samsum dataset [6]. The pre-trained model was further fine-tuned on the CANARD dataset. The objective of the fine-tuning process was to train the model with a sequence-to-sequence objective, enabling it to generate the current turn based on the history of the conversation up to that point.

The fine-tuning process was conducted using a distributed setup employing 2 V100 GPUs. The Hugging Face Trainer [7] was utilized for model training. Specifically, we set the gradient accumulation step to 2, with a per-device batch size of 32. A constant learning rate scheduler with warm-up steps was employed for training.

3.2 PTKB selection and ranking

The PTKB serves as a supplementary static context regarding the user, aiding in query expansion particularly when utterances exhibit ambiguity. Acting as an additional assumption, it assists the system in resolving ambiguity while responding to user queries. The selection of PTKB from the available options depends on its relevance to the current turn. Accordingly, the utterance of the current turn can be reformulated to better align with the selected PTKB.

Addressing this challenge as a similarity search problem between two sentences, we conducted similarity searches using embeddings generated by a language model. To accomplish this task, we employed the PCL pre-trained model (Wu et al., 2022).

4 Result and Discussion

Our team participated in the automatic track of the task, submitting a total of 4 runs. The results of these submissions are presented in Table-4. These results are reported on a non-pruned qrels dataset, which comprises 176 queries and 8716 relevant passages.

Furthermore, we conducted evaluations of our official submissions on a pruned qrels dataset. The pruned qrels dataset contains 133 queries and 5701 relevant passages. The results of this evaluation are presented in Table-3.

	dense_monot5	llm_damo
P@20	0.1831	0.1102
R@20	0.0812	0.0487
nDCG@3	0.2167	0.1343
nDCG@5	0.2206	0.1411
nDCG	0.2147	0.1105
mAP	0.0754	0.0376

Table 3: Performance of the system compared against pruned qrels

4.1 Performance against non-pruned Qrels

A thorough analysis was undertaken to evaluate the performance of the our best submission in terms of nDCG (normalized Discounted Cumulative Gain) metrics, in comparison to the median performance. nDCG, being a widely recognized metric for assessing the quality of ranked search results or recommendations, served as the primary evaluation criterion. The aim was to examine the degree to which the top-performing submission outperformed the median performance level.

The results of this analysis have been classified into two separate performance categories. The first category encompasses instances where the model's performance surpassed that of the median score. The second category pertains to instances where the model's performance only met the lower boundary of performance expectations.

Out of the 176 instances evaluated, the top-performing model demonstrated superiority over the median score in 90 instances. In these 90 instances, the nDCG score exceeded 0.0, indicating the model's ability to deliver search results or recommendations of enhanced relevance and quality in those particular scenarios.

[6] https://huggingface.co/datasets/samsum

[7] https://huggingface.co/docs/transformers/main_classes/trainer

Run Name	P@5	nDCG@5
run_automatic_dense_monot5	0.5634	0.5102
run_automatic_dense_damo_canard_16000	0.6512	0.6066
run_automatic_llm_damo	0.5247	0.4748
run_automatic_dense_mini_LM_reranker	0.6527	0.6225

Table 4: Performance of the system in PTKB ranking task

4.2 Performance against pruned Qrels

In the absence of specific information regarding the performance of other systems against pruned Qrels, we made the assumption that their performance would have been similar to their performance against non-pruned Qrels. With this assumption in mind, we conducted an analysis of the results, focusing on instances where our top-performing model either failed to return any relevant passages or retrieved substantially fewer passages (less than 10% of the relevant passages).

The core retrieval step, serving as the primary stage, heavily relies on the reformulated query generated by the query reformulator. This process is particularly sensitive to the keywords derived from the preceding conversation turn. A detailed analysis was conducted on a specific turn labeled '10-3_5' due to the presence of 33 relevant passages in Qrels, while our system failed to retrieve any passage within the top 1000 results.

Upon investigation, a thorough examination of the query reformulation process revealed a notable prevalence of anchor keywords primarily centered around the domain of food. Despite the absence of the term 'Italian' from both the Personal Textual Knowledge Base (PTKB) and the conversation flow, the existence of other keywords steered the query in an erroneous direction.

Furthermore, our analysis extended to the examination of other conversational turns, where analogous issues were identified within the query reformulation phase. It is noteworthy that a recurring pattern of these issues was observed in subsequent steps as well. Additionally, instances were noted where the prominence of extracted keywords from previous turns exerted a disproportionately influential effect, thereby disrupting the natural flow of communication and diverting it in unintended directions.

4.3 PTKB ranking

We utilized the PCL-bert model [8] to collect embeddings, and subsequently generated a similarity score between the utterance and the Personal Textual Knowledge Base (PTKB) corpus. Both raw and reformulated queries were employed for PTKB selection, with the top-3 relevant statements being selected based on the relevance score. The results were returned in an ordered manner, sorted by relevance score. The performance of the system across different runs is presented in Table-4.

Upon analysis, it was observed that the reformulated query-based PTKB selection yielded better results. However, the utilization of an expanded query with more verbose information resulted in subpar retrieval performance. This outcome may be attributed to the presence of excessive anchor information text in the reformulated query, which potentially introduced noise and hindered the retrieval process.

Conclusion

The paper presents a multi-step pipeline incorporating an intermediate step aimed at preserving conversational context. The initial stage of the pipeline utilizes query reformulation to enhance recall. However, query reformulation presents a potential drawback, as overly verbose queries may diminish retrieval performance. Moreover, the prominence of context keywords can disrupt the conversational flow during the reformulation process.

While our best submission demonstrated superior performance compared to the median score in more than half of the instances on a turn-by-turn basis, the system requires enhancement in two critical areas: the generation of concise queries and the intelligent handling of context keywords prior to their utilization in query reformulation.

The consistent selection of relevant passages from the Personal Textual Knowledge Base (PTKB) has proven beneficial for the retrieval en-

[8] https://huggingface.co/qiyuw/pcl-bert-base-uncased

gine, particularly when the utterance lacks sufficient information to generate an effective query. However, it was observed that implementing a cutoff score for PTKB selection would have been advantageous. Additionally, the static nature of the selection process, whereby the top-3 relevant PTKB are consistently chosen, has led to inadequately reformulated queries.

Improvements are needed to facilitate the generation of concise queries during the query reformulation stage. Furthermore, the system should be capable of identifying and managing context keywords in a manner that does not disrupt the natural flow of conversation.

Acknowledgements

Authors acknowledge the support of DST INSPIRE, Department of Science and Technology, Government of India.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Shivani Choudhary. 2022. Iitd-dbai: Multi-stage retrieval with pseudo-relevance feedback and query reformulation. *ArXiv*, abs/2203.17042.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with visual and semantic information.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy J. Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *ArXiv*, abs/2101.05667.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12052–12066, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanzhao Zhang, Dingkun Long, Guangwei Xu, and Pengjun Xie. 2022. HLATR: enhance multi-stage text retrieval with hybrid list aware transformer reranking. *CoRR*, abs/2205.10569.