Informatics and Health 1 (2024) 93–99

Contents lists available at ScienceDirect

Informatics and Health

journal homepage: www.keaipublishing.com/en/journals/informatics-and-health/

Exploring a learning-to-rank approach to enhance the Retrieval Augmented Generation (RAG)-based electronic medical records search engines

Cheng Ye *

Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

ARTICLE INFO

Keywords:
Retrieval Augmented Generation
Electronic medical records
Information retrieval
Large Language Model
Learning to rank

ABSTRACT

Background: This study addresses the challenge of enhancing Retrieval Augmented Generation (RAG) search engines for electronic medical records (EMR) by learning users' distinct search semantics. The specific aim is to develop a learning-to-rank system that improves the accuracy and relevance of search results to support RAG-based search engines.

Methods: Given a prompt or search query, the system first asks the user to label a few randomly selected documents, which contain some keywords, as relevant to the prompt or not. The system then identifies relevant sentences and adjusts word similarities by updating a medical semantic embedding. New documents are ranked by the number of relevant sentences identified by the weighted embedding. Only the top-ranked documents and sentences are provided to a Large-Language-Model (LLM) to generate answers for further review.

Findings: To evaluate our approach, four medical researchers labeled documents based on their relevance to specific diseases. We measured the information retrieval performance of our approach and two baseline methods. Results show that our approach achieved at least a 0.60 Precision-at-10 (P @ 10) score with only ten positive labels, outperforming the baseline methods. In our pilot study, we demonstrate that the learned semantic preference can transfer to the analysis of unseen datasets, boosting the accuracy of an RAG model in extracting and explaining cancer progression diagnoses from 0.14 to 0.50.

Interpretation: This study demonstrates that a customized learning-to-rank method can enhance state-of-the-art natural language models, such as LLMs, by quickly adapting to users' semantics. This approach supports EMR document retrieval and helps RAG models generate clinically meaningful answers to specific questions, underscoring the potential of user-tailored learning-to-rank methods in clinical practice.

Introduction

Electronic medical records (EMRs) store diagnosis and treatment details and are, therefore, a potential resource for medical researchers.[1–3] Unfortunately, scrolling through vast amounts of unstructured, redundant clinical text to identify relevant notes is time-consuming and expensive.[4,5]

EMR search engines are efficient tools to support information retrieval from EMRs.[6–11] A technique named "learning-to-rank" was introduced to enhance EMR search engines. Learning-to-rank[12,13] re-ranks search results by learning from labels provided by users. In general, a learning-to-rank approach represents each document by a set of features, such as bag-of-words. It then trains a classification model, such as a support vector machine or logistic regression, with user-provided labels to re-rank the search result.[13,14] Learning-to-rank has been widely used in web search engines, such as Google,[15] and recommendation systems in online retail, such as Amazon.[16] In medical research, researchers also applied learning-to-rank approaches to identify important terms in a clinical document[17,18] or to re-rank clinical documents to support medical research.[19]

Searching across medical records and producing labels is challenging because of an interesting paradox: even if a document contains a search term, a user may determine the document is not relevant to the search. For example, a note may reference "diabetes" in the past medical history, but the primary purpose of the note is a recent leg injury. While traditional learning-to-rank approaches can help focus the search on a user's semantics, they can be limited because they require an offline training process and a large amount of training data, which is not suitable for one-time chart review tasks. Moreover, traditional learning-to-rank tasks that use bag-of-word features are not able to update the ranking

* Correspondence to: Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave # 1475, Nashville, TN 37203, USA.
E-mail address: cheng.ye.1@vumc.org.

https://doi.org/10.1016/j.infoh.2024.07.001
Received 14 March 2024; Received in revised form 8 July 2024; Accepted 9 July 2024
Available online 23 July 2024

for words that have not yet been seen by the system.

As the state-of-the-art natural language model, the Large-Language-Models (LLMs),[20] such as the GPT-3.5 and GPT-4,[21] Google's Gemini,[22] came to exist, one type of next-generation search engine, called Retrieval Augmented Generation (RAG)-based search engines, attract more and more attention, in both the artificial intelligence research field. An RAG-based search engine first retrieves relevant text segments from a target database based on the query or prompt provided by the users, and then the query and relevant text segments are provided to an LLM or multiple LLMs to generate answers to users' query or prompt. A RAG-based search engine can efficiently overcome the hallucination issues when using the LLMs and can be a critical component of a reliable generative AI, especially for the application of generative AI in healthcare, which has specific legislations.[23–26] Therefore, specific explorations of RAG-based search engines are needed better to shape the future research directions of next-generation generative AI and enhance the reliability and trustworthiness of the application of generative AI in healthcare.

In this paper, we explore a novel learning-to-rank approach that learns to rank in the early stage (i.e., before sending search results to RAG-based LLMs) of chart review tasks with a limited number of positive labels. Given a clinical chart review task, as shown in Fig. 1, the system first extracts primary keywords from the clinical questions of the task identifies sentences in a document that are relevant to the keywords, and their expanded semantic similar words based on semantic models, such as word embeddings,[27] medical semantic models.[28] Then, we train a learning-to-rank model that learns to weigh the relevant sentences, which contain the keywords and their semantic similar words. The system learns to re-rank the relevant sentences and, therefore, refine the overall ranks of documents by re-computing the number of re-ranked positive sentences minus re-ranked negative sentences in the documents. After training a learning-to-rank model to adjust to the given clinical chart review task and the specific semantic needs of the users, the system moves to Step 2, as shown in Fig. 1. The system first transfers the clinical research questions into the pre-designed prompt template and then retrieves documents based on the recommendation of the trained learning-to-rank model. A selected Large Language Model (e.g., the GPT 3.5) receives the prompts and the refined document subset as the inputs and generates human-friendly answers. Such a hybrid approach that integrates the traditional learning-to-rank approach and the state-of-the-art (SOTA) LLM is able to generate explainable and reliable answers to support clinical chart reviews in a rapid and accurate manner.

In this study, to evaluate the proposed learning-to-rank approach, we selected two finished clinical chart review tasks and extracted training and evaluation datasets from their action log files. In each of these two clinical chart review tasks, four medical researchers labeled a set of clinical documents based on a document's relevance to the primary keywords of the clinical chart review tasks: "diabetes" or "seizure":

(1) Clinical Chart review Task 1. In this task, four medical researchers reviewed all 300 clinical notes and selected sentences that are related to barriers of diabetes plans. To construct the training and evaluation dataset, for this task, the primary keyword is "diabetes." "In this study, we select the pre-trained EMR-based word embeddings to expand the query with the similar words of "diabetes", based on our previous work, the EMR-subsets method.[27]
(2) Clinical Chart Review Task 2. In this task, four medical researchers reviewed all 300 clinical notes and selected sentences that are related to the diagnosis and treatment of "seizure." For this task, we applied the same approach mentioned above to construct the training and evaluation datasets.

With these two datasets, the goals were to evaluate that we can learn to capture the semantic needs of users with a small set of positive sentences since a learning-to-rank model that requires a large amount of training labels is not a practical solution in real-world clinical chart reviews. In each of these two evaluation experiments, the learning-to-rank approach was first trained with ten positive samples and then evaluated for its average Precision-at-10 (P @ 10) score using ten-fold cross-validation. The results show our approach achieved higher P @ 10 in more of the evaluations than the baseline ranking methods (e.g., 0.60 P @ 10 vs. 0.40 P @ 10).

After we confirmed that the proposed learning-to-rank approach is practical, we then evaluated the learning-to-rank approach in another clinical chart review task, with a full evaluation of Steps 1, 2, and 3 shown in Fig. 1. The evaluation result shows that, without a learning-to-rank approach, the average F1 score of the answer provided by the LLMs is around 0.22 in answering cancer-stage questions, while a learning-to-rank approach boosted the average F1 score to around 0.68. Such an improvement is due to the quality of the refined search results provided by the trained learning-to-rank approach.

In the rest of this paper, we first introduce the methods we used in this study in more detail and then present and discuss the evaluation results of our proposed approach.

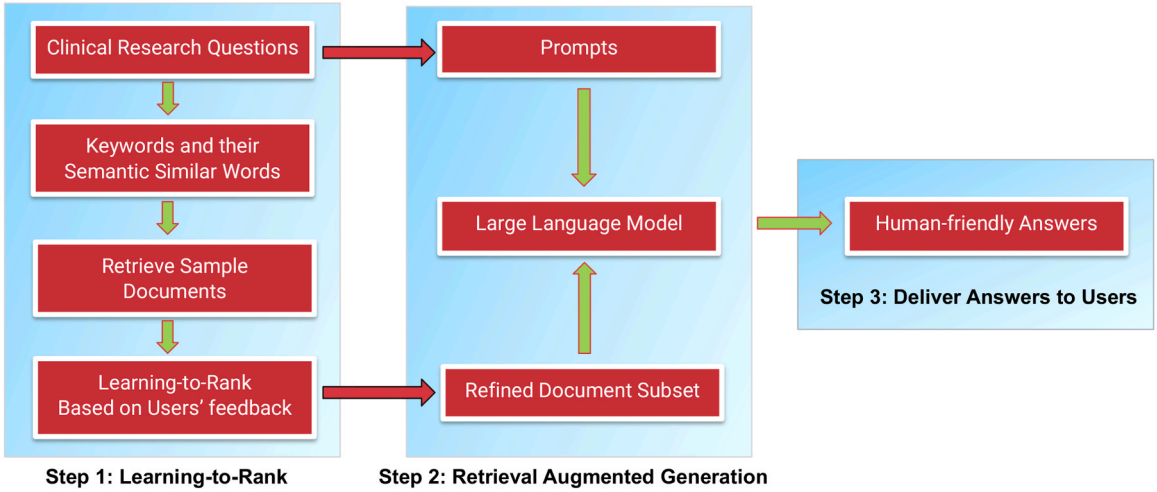

Fig. 1. The learning-to-rank approach to support Retrieval Augmented Generation (RAG) during clinical chart reviews.

Methods

Definitions

A user determines if a document is relevant to a search by considering the existence of keywords in the document and the words surrounding the keywords. These words surrounding the keywords can be used to infer the semantics of the search beyond a single search term. Therefore, as users identify which documents are relevant or not, the system can learn the user's search semantics.

Let a **relevant sentence** be a sentence that contains at least one **relevant word** to the search term (e.g., "metformin" is relevant to "diabetes"). Each word in the relevant sentence has an **impact value**. A relevant sentence is considered to be **positive** if the sum of impact values for all words in the sentence is positive and negative otherwise. Let a **positive document** be a document that contains more positive sentences than negative sentences and a **negative document** be one that contains more negative sentences than positive sentences.

Semantic embedding

To expand the search query for retrieving documents during a clinical chart review task, researchers have their own preferences for selecting semantic embeddings to generate expansion words to enhance the keywords within the search query. In this study, we introduce a simple word embedding, other than introducing other more mature semantic approaches (e.g., the EMR-subsets method[29] and the medical context vector space [30]), to demonstrate the generalizability of the proposed learning-to-rank approach. In future work, we plan to expand the evaluation of our study with more state-of-the-art semantic embeddings or LLMs.

Given a search term, the relevant sentences in a document are determined using a word embedding trained with clinical notes from the Vanderbilt University Medical Center Synthetic Derivative,[31] a de-identified mirror of our EMR, which contained approximately 100 million clinical notes at the time of this study.

A word embedding projects words into a vector space by training a neural network with text[32,33]. We used the positions of words in the embedded vector space to estimate their similarity. We measured the similarity of words w_i and w_j using the cosine similarity of their embedded vectors v_i and v_j. The range of similarity is from zero to one.

$$s(w_i, w_j) = \frac{v_i \bullet v_j}{\| v_i \| \times \| v_j \|}$$

When the ranking algorithm encounters unseen words in extracted relevant sentences during the prediction, we estimated the impact value of an unseen word with the average impact of all of its similar words in the training set.

A word w is **relevant** to the search term t if its similarity is larger than a pre-defined similarity **cutoff**. We define the **impact** of a word w to search term t as a weighted similarity that fits a user's semantics.

$$\text{Impact}(w, t) = c_w * s(w, t)$$

We define the impact of a **relevant sentence** S_i to search term t as follows.

$$\text{Impact}(S_i, t) = \sum_{w \in S_i} \text{Impact}(w, t)$$

A **positive document** is a document D that has a positive impact on the search term t.

$$\text{Relevance}(D, t) = \begin{cases} 1, \text{if} \sum_{S_i \in D} \text{Impact}(S_i, t) > 0 \\ 0, \text{if} \sum_{S_i \in D} \text{Impact}(S_i, t) \le 0 \end{cases}$$

Learning the impact values

Consider a set of positive-labeled documents D_p and a set of negative-labeled documents D_n for a search term. The learning goal is to identify a set of words with their impact values and corresponding coefficients (i.e., an impact set I) to minimize the loss function as follows:

$$\min_I \sum_{D_i \in (D_p \cup D_n)} |\text{Relevance}(D_i, t) - \text{Label}(D_i, t)|$$

We first extract all relevant sentences similar to the search term t using the word embedding, excluding unrelated sentences. We then chose a **similarity cutoff C** for the sentences in the document to maximize the total similarity of selected sentences to the search term and minimize the number of selected sentences.

$$\max_C \frac{||S||}{N} * \sum_{S_i \in S} (s(S_i, t) - C)$$

In each iteration of the learning process, for each positive sentence P_i, its absolute impact should be larger than the absolute impact of a negative sentence N_j. This condition is based on the observation that a user is more likely to label a document as relevant when given one positive sentence and one negative sentence if the positive sentence has a larger impact.

$$\begin{cases} \text{Impact}(P_i, t) > 0 \\ \text{Impact}(N_j, t) \le 0 \\ |\text{Impact}(P_i, t)| > |\text{Impact}(N_j, t)| \end{cases}$$

We initialize the relevant sentences in a positive-labeled document to be positive and all relevant sentences in a negative-labeled document to be negative. We set the initial impact values of words as their similarities to the search term t based on the word embedding.

For each document in the test dataset, if there was no relevant sentence, it was excluded from the search result. Otherwise, we first compute the impact values of relevant sentences by summing up the impact values of words in the sentences and then count the number of positive and negative sentences to determine the document's impact. For a word w′ that had no impact record in the training data, we selected its most similar word using the word embedding and used its most similar word's impact value.

Finally, we rank all unlabeled documents by the number of positive sentences minus negative sentences in each document.

Evaluation

We randomly identified 300 notes from patients in the EMR system with an ICD-9 code for "diabetes" (250.*) and another 300 notes from patients with an ICD-9 code for "seizure" (780.39). The details of the selected notes are:

(1) For clinical chart review task 1, we followed the suggestions of the participating medical researchers to randomly select 300 progression notes from patients with ICD-9 code (250. *) and with at least one mention of "diabetes" in at least one of their notes within the first month of 2016 in the EMR database. The selected notes have an average length of 1450 words per note; 97 % of the notes contain the keyword "diabetes."
(2) For clinical chart review task 2, we followed the suggestions of the participating medical researchers to randomly select 300 diagnosis notes from patients with ICD-9 code (780.39) and with at least one mention of "seizure" in at least one of their notes within the first month of 2016 in the EMR database. The selected notes have an average length of 1057 words per note; 22.3 % of the notes contain the keyword "seizure."

In each of the clinical review tasks, we asked four medical

researchers (referred to as users 1, 2, 3, 4) to label each note's relevance to a disease (1-relevant or partially relevant, or 0-irrelevant). Relevant notes were considered positive, while irrelevant notes were considered negative. We received four labeled document sets for the "diabetes" cohort (Table 1) and four labeled document sets for the "seizure" cohort (Table 2). We noticed that different medical researchers labeled different positive notes and, therefore, revealed that different medical researchers have different labeling strategies. In one of our previous works,[34] we extensively analyzed such a difference in labeling important notes. Due to the redundancy of EMRs, there exist multiple subsets of clinical notes in a given clinical chart review task that provide the same conclusion to the research questions of the task. For example, as shown in Table 1, User 3 prefers to first rank the notes by date and only review the latest notes of a patient, while User 1 prefers to review all notes and label all positive notes. We refer the reader to our previous work [34] for more details about the difference in document preference across medical researchers.

For comparison, we selected two baseline ranking methods, which are commonly used in EMR search engines: rank by the number of search terms in the documents and by the number of similar words of the search terms in the documents.

We tested the ranking methods in two application scenarios:

(1) For each document set, we shuffled it and selected the first K documents, which included ten positive labeled documents. We selected K with the elbow method proposed in our previous work.[27] The elbow method is able to identify an optimal point to select a high-quality subset of data points from a large dataset. We first computed the overall similarity of each document to the keywords of the search query and then generated the similarity curve to compute the K value as the "elbow" point of the similarity curve. Then, we trained our learning-to-rank algorithm and measured the Precision-at-10 (P @ 10) score of all ranking methods using the other (300 - K) documents. We repeated the test 10 times and measured the average P @ 10 score of each ranking method.

(2) For each document set, we randomly shuffled it, selected the first 50 % of documents, and then trained our learning-to-rank algorithm. After that, we measured the average P @ 10 score of each ranking method using the other 50 % documents. We repeated the test 10 times and measured the average P @ 10 score of each ranking method.

Results

The evaluation results show that our learning-to-rank approach outperformed baseline methods in all of the labeled document sets for the "diabetes" cohort. As shown in Table 3, with ten positive diabetes labels provided by users 3 and 4, our learning-to-rank approach achieved much higher average P @ 10 scores than the baseline methods. The P @ 10 scores from learning-to-rank were twice those from the similar word ranking approach (0.60 vs. 0.30 for user 3 and 0.80 vs. 0.40 for user 4). Labels provided by users 1 and 2 also generated improved P @ 10 scores using learning-to-rank over both baseline rankings, but these effects were not as robust.

When we included 50 % of the documents in our diabetes dataset (Table 4), we observed even stronger effects. Our learning-to-rank approach achieved at least a 0.80 P @ 10 score in the labeled document sets provided by all users (Table 4). The performance of our learning-to-rank approach increased when more labels were given to most users. Especially for user 3, the P @ 10 score increased from 0.60 (Table 3) to 0.80 (Table 4).

As shown in Table 5, our learning-to-rank approach achieved minimal gains over the baseline methods when trained with ten positive seizure labels. However, when more training data was included (by including 50 % of the documents in our seizure dataset), the performance of the learning-to-rank approach increased and outperformed the baseline methods, as shown in Table 6. For all users, our learning-to-rank approach achieved at least a 0.90 P @ 10 score, whereas the ranking by number of search terms and ranking by number of similar words achieved a P @ 10 score as low as 0.60 (Table 6, user 4). Especially for user 3, when given more labels, the P @ 10 increased from 0.85 (Table 5) to 0.90 (Table 6).

Tables 7 and 8 present the words with the highest positive/negative impact on user decisions when searching "diabetes." As shown in Table 7, words related to medications, treatments, or problems had a positive impact. As shown in Table 8, words about family or medical history had a negative impact on users' labels.

Users had both common and unique words that impacted their decisions. For example, "health," "problem," "father," and "fasting" were common words that impact users' decisions (Table 8), whereas "lab," "knee," and "post" were user-specific.

Table 1
Distribution of positive and negative notes of the evaluation datasets for "diabetes.".

User	Positive notes	Negative notes
1	103	197
2	109	191
3	24	276
4	63	237

Table 2
Distribution of positive and negative notes of the evaluation datasets for "seizure.".

User	Positive notes	Negative notes
1	103	210
2	109	156
3	26	274
4	33	267

Table 3
Average P @ 10 scores when searching "diabetes." For each user, the learning-to-rank approach was trained with K documents, including ten positive notes.

P @ 10 Evaluation	User 1	User 2	User 3	User 4
Learning to Rank	**0.95**	**0.90**	**0.60**	**0.80**
Rank by number of search terms	0.85	0.85	0.40	0.70
Rank by number of similar words	0.85	0.80	0.30	0.40

Table 4
Average P @ 10 scores when searching "diabetes." For each user, the learning-to-rank approach was trained with 50 % of the labeled documents.

P @ 10 Evaluation	User 1	User 2	User 3	User 4
Learning to Rank	**0.95**	**0.95**	**0.80**	**0.90**
Rank by number of search terms	0.85	0.85	0.40	0.70
Rank by number of similar words	0.85	0.80	0.30	0.40

Table 5
Average P @ 10 scores when searching "seizure." For each user, the learning-to-rank approach was trained with K documents, including ten positive notes.

P @ 10 Evaluation	User 1	User 2	User 3	User 4
Learning to Rank	0.95	**0.95**	0.85	0.90
Rank by number of keywords	0.95	0.90	0.85	0.90
Rank by number of similar words	0.85	0.80	0.65	0.60

Table 6
Average P @ 10 scores when searching "seizure." For each user, the learning-to-rank approach was trained with 50 % of the labeled documents.

P @ 10 Evaluation	User 1	User 2	User 3	User 4
Learning to Rank	0.95	**0.95**	**0.90**	**0.95**
Rank by number of keywords	0.95	0.90	0.85	0.90
Rank by number of similar words	0.85	0.80	0.65	0.60

Table 7
Top positive words for users when searching "diabetes.".

Index	User 1	User 2	User 3	User 4
1	health	health	identifying	metformin
2	problems	tablet	identified	assessment
3	mouth	problems	increase	increase
4	daily	mouth	problems	Plan
5	mg	daily	order	dm

Table 8
Top negative words for users when searching "diabetes.".

Index	User 1	User 2	User 3	User 4
1	cancer	diverticular	father	Knee
2	father	lab	past	replacement
3	family	family	history	Pain
4	smoking	schedule	post	Health
5	fasting	fasting	children	medication

Discussion

Although learning-to-rank has been studied thoroughly in many research areas, little research has been done on learning from a limited number of positive-labeled documents in one-time chart review environments. Here, we provide evidence that our learning-to-rank approach offers an advantage over other ranking methods and can be generalized to unseen data after being trained with a limited EMR dataset.

Our approach achieved high average P @ 10 scores across users with a limited number of positive-labeled documents, and the P @ 10 score further increased when more training data was added. This suggests the approach may be of particular benefit in the early stages of a chart review task when few labels are available. Our approach is able to learn and support users with a limited number of labels (e.g., in Table 1, user 3 only labeled 24 positive notes). As users provide more labels, the performance will continue to increase, as shown in Tables 4 and 6. For users looking to perform a single chart review, this approach can be used with minimal training requirements.

The most difficult parts of utilizing limited positive samples (e.g., less than 10) are how to extract enough high-quality training data for a machine-learning model and how to generalize a learned model to unseen data. In our study, we introduced an assumption that a positive sentence always has a larger impact than a negative sentence. This assumption is necessary because (1) When given a new document with only one positive and one negative sentence, we always assume that a user is more likely to label it as a positive document; (2) It generates more effective training data to learn the impact values because we can consider all pairs of positive and negative sentences. For example, with ten positive sentences and ten negative sentences, we have 100 training samples. Similar to our assumption for building a training data set, the SVM-ranking[13] method extracts users' preferences from clickthrough data. However, our approach addresses a critical application scenario. Our approach requires few labels, while other learning-to-rank approaches like SVM need hundreds of training examples. Therefore, we paid specific attention to over-generalization by introducing the sentence preference conditions (i.e., a positive sentence has a larger impact than a negative sentence) and handled unseen data by introducing word embedding trained with EMRs.

Our study also used only words as features for a learning-to-rank model and achieved higher average P @ 10 scores than other methods. Although adding non-language features (e.g., the length of documents, the type of documents) may improve rankings, it is hard to present and explain those features to users. Our approach is able to re-rank search results close to users' preferences and explain the ranking criteria to users by highlighting the words that impact their decisions.

A word embedding is critical when generalizing the impacts of words from the training data to unseen data. For example, if we identify "father" and "family" as negative words for a user (Table 8), then the words "sister" and "brother" may also have negative impacts on this user's decision. Therefore, we could generalize the impacts of the family-related words from the training data to unseen data.

Interestingly, ranking the result by the number of similar words achieved the lowest P @ 10 score in all tests. As shown in Table 3, user 4, ranking by the number of similar words, achieved a 0.40 P @ 10 score, while ranking by the number of keywords, achieved a 0.70 P @ 10 score. We expected that ranking by the number of similar words would perform better than ranking by the number of search terms since it considers relevant information of the query. Therefore, this demonstrates that one should be careful about using the raw similarities to provide a word embedding in a natural language processing task. There are some limitations and possible future work of this study. We assumed only the words surrounding the keywords of a search term or the similar words of the search term impact the decision of users. However, words or sentences that are far away from the keywords in a document may also impact a user's decision. Future studies may require more chart review tasks, such as asking users to provide both labels and snippets of text that impact their decisions. The addition of snippets would provide a better understanding of how medical researchers read documents and improve our learning-to-rank method. To simplify the study, we asked users to provide binary labels of relevance; however, the labels could be extended to a larger range (e.g., 0–10) in the future. In that case, we may further analyze users' preferences for documents and improve our learning-to-rank approach.

Also, in this paper, we only introduced the basic word embedding structure to demonstrate the generalizability of our learning-to-rank approach. As the next step, we may explore, introduce, and evaluate some of the most recent, state-of-the-art approaches in building EMR-related models to support labeling in different grain-level, such as in entity labeling,[35] sequence/sentence level labeling,[36] and document level labeling.[37] We also plan to explore our research on capturing users' semantic preferences from the traditional learning-to-rank approach to the few-shot, deep learning-based approach.[38]

Finally, at the time of submitting this paper, we only measured the P @ 10 score of our approach. In the future, we plan to do more information retrieval analyses, such as measuring P @ K, F1 score, and recall. Moreover, we plan to test the proposed learning-to-rank approach with some state-of-the-art LLMs (e.g., GPT3.5, GPT 4) or open-source RAG framework (e.g., the llama index [39]). In one of our pilot studies, we evaluated the whole framework shown in Fig. 1, with a dataset extracted from a clinical chart review task that focused on the labeling of cancer-progression-related sentences and documents within progression notes. Based on the evaluation analysis of the pilot study, we found that (1) the learned semantic preference can also be transferred to the analysis of the unseen dataset and boosted the average accuracy of the RAG model in extracting and explaining the cancer progressing diagnosis from 0.14 to 0.50; and (2) LLMs still need the support of retrieval augmentation. Without the support of the expanded query (i.e., expanding the keywords with their similar words), the LLMs perform poorly in the clinical chart review task 1 of this study. One of the key reasons is most of the relevant sentences do not contain the keyword "diabetes" but contain the similar words of "diabetes.".

As the next step, we plan to test more SOTA LLMs and evaluate and

compare the performance of our learning-to-rank approach when working with different types of LLMs. In this study, we only tested the framework shown in Fig. 1 with GPT3.5 and Llama 2 70B model [39] to extract and explain the answer to the cancer progression of a patient.

Conclusion

In this study, we proposed a learning-to-rank approach that quickly learns a user's search semantics from a limited number of labeled documents. To demonstrate the generalizability of our approach, we selected a simple word embedding to help identify sentences and words that may impact a user's decision and adjust the weights of words to learn a user's search semantics. Moreover, our approach can be easily generalized to measure the impact values of unseen words using word embeddings or other semantic approaches (e.g., the BERT and BioBERT [30]). The evaluation shows that, by learning the impact of words and sentences on the search term, our approach achieved better Precision at 10 (P @ 10) scores than baseline methods when only ten positive-labeled documents were provided. The learning-to-rank method may be used to support retrieving documents during a chart review with minimal training requirements and continues to improve its quality through active learning or be used as the foundation of a Retrieval Augmented Generation (RAG)-based generative AI applied in healthcare industry.

Funding

Crowd Sourcing Labels from Electronic Medical Records to Enable Biomedical Research Award Number: 1 UH2 CA203708-01.

CRediT authorship contribution statement

Cheng Ye: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors whose name are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript:

Acknowledgments

The training data for the word2vec embeddings was obtained from VUMC's Synthetic Derivative, which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH.

Competing interests

None.

Provenance and peer review

Not commissioned; externally peer-reviewed.

References

1. Rasmussen LV. The electronic health record for translational research. *J Cardiovasc Transl Res*. 2014;vol. 7(6):607–614. https://doi.org/10.1007/s12265-014-9579-z.
2. Chen L, et al. Racing against the clock: internal medicine residents' time spent on electronic health records. *J Grad Med Educ*. 2016;vol. 8(1):39–44. https://doi.org/10.4300/JGME-D-15-00240.1.
3. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. *J Am Med Inf Assoc*. 2011;vol. 18 (2):112–117. https://doi.org/10.1136/jamia.2010.008441.
4. Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. *J Am Med Inform Assoc*. 2010;vol. 17(1): 49–53. https://doi.org/10.1197/jamia.M3390.
5. Natarajan K. *Analysis of Search on Clinical Narrative within the EHR*. Columbia University; 2012.
6. Biron P, Metzger MH, Pezet C, Sebban C, Barthuet E, Durand T. An information retrieval system for computerized patient records in the context of a daily hospital practice: the example of the Léon Bérard Cancer Center (France). *Appl Clin Inf*. 2014; vol. 5(1):191–205. https://doi.org/10.4338/ACI-2013-08-CR-0065.
7. Natarajan K, Stein D, Jain S, Elhadad N. An analysis of clinical queries in an electronic health record search utility. *Int J Med Inf*. Jul. 2010;vol. 79(7):515–522. https://doi.org/10.1016/j.ijmedinf.2010.03.004.
8. Tawfik AA, Kochendorfer KM, Saparova D, Al Ghenaimi S, Moore JL. I don't have time to dig back through this': The role of semantic search in supporting physician information seeking in an electronic health record. *Perform Improv Q*. 2014;vol. 26 (4):75–91. https://doi.org/10.1002/piq.21158.
9. Zalis M, Harris M. *Advanced search of the electronic medical record: Augmenting safety and efficiency in radiology*. In: *Journal of the American College of Radiology*. vol. 7. Elsevier; Aug. 2010:625–633. https://doi.org/10.1016/j.jacr.2010.03.011.
10. Gregg W, Jirjis J, Lorenzi NM, Giuse D. StarTracker: an integrated, web-based clinical search engine. *AMIA ... Annu Symp Proc / AMIA Symp AMIA Symp*. 2003;vol. 2003(1):855.
11. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: a report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *J Biomed Inf*. 2015;vol. 55:290–300. https://doi.org/10.1016/j.jbi.2015.05.003.
12. Nakamoto Y. A Short Introduction to Learning to Rank. *IEICE Trans Inf Syst*. 2011; vol. E94-D(1):1–2. https://doi.org/10.1587/transinf.E94.D.1.
13. T. Joachims, "Optimizing search engines using clickthrough data", *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, p. 133, 2002, ⟨doi:10.1145/775066.775067⟩.
14. B. Li, R. Xiao, Z. Li, R. Cai, B.L. Lu, and L. Zhang, "Rank-SIFT: Learning to rank repeatable local interest points", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1737–1744, 2011, doi: 10.1109/CVPR.2011.5995461.
15. D. Sculley, "Large Scale Learning to Rank", *NIPS 2009 Workshop on Advances in Ranking*, pp. 1–6, 2009.
16. D. Sorokina and E. Cantú-paz, "Amazon Search: The Joy of Ranking Products", *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*, pp. 459–460, 2016, ⟨doi:10.1145/2911451.2926725⟩.
17. J. Chen, A.N. Jagannatha, S.J. Jarad, and H. Yu, "Ranking medical jargon in electronic health record notes by adapted distant supervision", *arXiv preprint arXiv: 1611.04491*, 2016.
18. Chen J, Yu H. Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients. *J Biomed Inf*. 2017;vol. 68:121–131. https://doi.org/10.1016/j.jbi.2017.02.016.
19. Jin M, Li H, Schmid CH, Wallace BC. Using electronic medical records and physician data to improve information retrieval for evidence-based care. *IEEE Int Conf Healthc Inform (ICHI)*. 2016. https://doi.org/10.1109/ICHI.2016.12.
20. Chang Y, et al. A survey on evaluation of large language models. *ACM Trans Intell Syst Technol*. Jan. 2024. https://doi.org/10.1145/3641289.
21. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Sci Rep*. 2023;vol. 13(1), 20512. https://doi.org/10.1038/s41598-023-46995-z.
22. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. *Eye*. 2024:1–6.
23. Oniani D, et al. Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digit Med*. 2023;vol. 6(1):225.
24. Duffourc M, Gerke S. Generative AI in health care and liability risks for physicians and safety concerns for patients. *JAMA*. 2023.
25. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;vol. 6(1):120.
26. Ye C, et al. A crowdsourcing framework for medical data sets. *AMIA Summits Transl Sci Proc*. 2018;vol. 2017:273–280.
27. Ye C, Fabbri D. Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. *J Biomed Inf*. 2018;vol. 83(April). https://doi.org/10.1016/j.jbi.2018.05.014.
28. Ye C, Malin BA, Fabbri D. Leveraging medical context to recommend semantically similar terms for chart reviews. *BMC Med Inf Decis Mak*. 2021;vol. 21(1):353. https://doi.org/10.1186/s12911-021-01724-2.
29. Ye C, Fabbri D. Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews (p.) *J Biomed Inf*. 2018;vol. 83(April). https://doi.org/10.1016/j.jbi.2018.05.014.
30. Ye C, Malin BA, Fabbri D. Leveraging medical context to recommend semantically similar terms for chart reviews. *BMC Med Inf Decis Mak*. 2021;vol. 21(1):353. https://doi.org/10.1186/s12911-021-01724-2.
31. Roden DM, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharm Ther*. 2008;vol. 84(3):363.

32. T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space", *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12, 2013, ⟨doi:10.1162/153244303322533223⟩.
33. Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents", *International Conference on Machine Learning - ICML 2014*, vol. 32, pp. 1188–1196, 2014, ⟨doi:10.1145/2740908.2742760⟩.
34. Ye C, Fabbri D. Next generation of electronic medical record search engines to support chart reviews: a systematic user study and future research direction. *J Econ Technol*. 2024;vol. 2:22–30. https://doi.org/10.1016/j.ject.2024.03.003.
35. Li J, Shang S, Chen L. Domain generalization for named entity boundary detection via metalearning. *IEEE Trans Neural Netw Learn Syst*. 2021;vol. 32(9):3819–3830. https://doi.org/10.1109/TNNLS.2020.3015912.
36. Li J, Han P, Ren X, Hu J, Chen L, Shang S. Sequence labeling with meta-learning. *IEEE Trans Knowl Data Eng*. 2023;vol. 35(3):3072–3086. https://doi.org/10.1109/TKDE.2021.3118469.
37. J. Li, Y. Wang, S. Zhang, and M. Zhang, "Rethinking Document-Level Relation Extraction: A Reality Check", *ArXiv*, vol. abs/2306.08953, 2023, [Online]. Available: ⟨https://api.semanticscholar.org/CorpusID:259164990⟩.
38. Li J, Feng S, Chiu B. Few-shot relation extraction with dual graph neural network interaction. *IEEE Trans Neural Netw Learn Syst*. 2023:1–13. https://doi.org/10.1109/TNNLS.2023.3278938.
39. J. Pinheiro *et al.*, "On the Construction of Database Interfaces Based on Large Language Models", 2023.