

Re²Dan: Retrieval of Medical Documents for e-Health in Danish

Antonela Tommasel

atommasel@cs.au.dk

Aarhus University

Department of Computer Science

DIGIT Aarhus University Centre for

Digitalisation, Big Data and Data Analytics
Denmark

ISISTAN, CONICET-UNCPSA, Argentina

Rafael Pablos Sarabia

rpablos@cs.au.dk

Aarhus University

Department of Computer Science

DIGIT Aarhus University Centre for

Digitalisation, Big Data and Data Analytics
Denmark

Ira Assent

ira@cs.au.dk

Aarhus University

Department of Computer Science

DIGIT Aarhus University Centre for

Digitalisation, Big Data and Data Analytics
Denmark

ABSTRACT

With the clinical environment becoming more data-reliant, healthcare professionals now have unparalleled access to comprehensive clinical information from numerous sources. Then, one of the main issues is how to avoid overloading practitioners with large amounts of (irrelevant) information while guiding them to the relevant documents for specific patient cases. Additional challenges appear due to the shortness of queries and the presence of long (and maybe noisy) contextual information. This demo presents Re²Dan, a web Retrieval and recommender of Danish medical documents. Re²Dan leverages several techniques to improve the quality of retrieved documents. First, it combines lexical and semantic searches to understand the meaning and context of user queries, allowing the retrieval of documents that are conceptually similar to the user's query. Second, it recommends similar queries, allowing users to discover related documents and insights. Third, when given contextual information (e.g., from patients' clinical notes), it suggests medical concepts to expand the user query, enabling a more focused search scope and thus obtaining more accurate recommendations. Preliminary analyses showed the effectiveness of the recommender in improving the relevance and comprehensiveness of recommendations, thereby assisting healthcare professionals in finding relevant information for informed decision-making.

CCS CONCEPTS

• Information systems → Recommender systems; Information retrieval query processing; Language models.

KEYWORDS

health care, document recommendation, natural language processing, medical documents

ACM Reference Format:

Antonela Tommasel, Rafael Pablos Sarabia, and Ira Assent. 2023. Re²Dan: Retrieval of Medical Documents for e-Health in Danish. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23)*, September 18–22, 2023, Singapore, Singapore. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3604915.3610655>

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).

RecSys '23, September 18–22, 2023, Singapore, Singapore

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0241-9/23/09.

<https://doi.org/10.1145/3604915.3610655>

1 INTRODUCTION

The growing availability of medical documents and the rise of electronic medical records have created a greater demand for specialized document retrieval and recommendation systems [6]. In recent years, the semantic medical search problem has been tackled by different scientific communities, perspectives, and methods [19]. The main challenge is understanding the complex information needs related to medical applications and developing methods to increase the likelihood of finding relevant documents [19]. Searching through medical documents poses specific challenges, especially when dealing with languages other than English, like Danish. For example, a particular symptom or a result in a laboratory test might lead a clinician to infer a specific diagnosis, even if it is not explicitly mentioned [8].

Often, practitioners lack time to develop effective search strategies, thus frequently retrieving irrelevant documents [15]. Another concern is related to how to design medical retrieval and clinical decision support systems that can effectively assist different practitioners' profiles, from novices to experts, in tackling complex tasks like searching for protocols, diagnosis, or treatments, all encompassed within the same system, and each with potentially different relevance requirements [19]. These different profiles will likely show differences in domain knowledge and expertise, significantly impacting the used vocabulary, query formulation, and document relevance assessment [16, 18, 19]. For example, Tamine and Chouquet [17] determined that experts tend to search for longer queries that include more technical concepts than novices, meaning that documents' rankings could significantly vary for similar information needs. In this context, due mainly to the different search strategies adopted by users and how document retrieval is implemented, different studies [15] have shown that general-purpose retrieval systems (such as Google or even Google Scholar) might be preferred over specialized systems (such as PubMed or MedLine).

In Denmark, practitioners have access to official clinical guidelines, medical documents, and work protocols through dedicated search interfaces. Each region has its own documents and web systems, which might significantly differ in appearance and usage [2]. For example, the eDok system¹ for the Central Denmark Region (Region Midtjylland) still relies only on simple lexical and synonyms search², which hinders the retrieval of semantically related documents or even the search of compound concepts.

To address these challenges, we need tools that can bridge the semantic gap between document descriptions and the conceptual

¹<https://e-dok.rm.dk/>

²<https://www.defactum.dk/om-DEFACTUM/kernejdelsel/forbedringsarbejde/metodekatalog/vidensdeling-og-spredning/e-dok/>

meaning of queries [15]. In this context, we present Re²Dan, a publicly available web system for Danish medical document retrieval and recommendation based on combining lexical and semantic search capabilities with similar query recommendation. To our knowledge, this is the first available recommender system based on semantic search for medical documents in Danish.

2 RE²DAN: OVERVIEW

Figure 1 presents the Re²Dan³ pipeline. A typical use case of Re²Dan consists of a user (mostly medical personnel) who writes a query and optionally adds some context to it, for example, from patients' clinical notes. The query is spell-checked, and similar queries are retrieved to be recommended to the user. If context is provided, medical concepts are highlighted, and the user can choose to add them to the query. Then, documents are ranked and shown to the user. Finally, queries and clicks are stored to be added to the pool of historical queries. We detail each component as follows.

Document representation. eDok is the public system for retrieving the guidelines, procedures, and medical documents belonging to the Center Region in Denmark. Documents follow a specific template with some predetermined headings. For representing documents, we considered their title, keywords, area, and department for whom the document can be accessed (to filter candidate results) and the text for each section.

Two representations were considered to support both lexical and semantic retrieval of documents. First, an inverted index representation (implemented using PyTerrier⁴), including word frequency and position. To reduce the lexical variations and improve similarity matching, words were lemmatized using spaCy⁵. Second, a semantic embedding representation to overcome the potential mismatch between query and document vocabularies. Given the limited availability of Danish models and the lack of specialized medical embeddings suitable for similarity tasks, after preliminary analyses, we opted for the Teacher-Student knowledge distillation approach [1] to adapt SentenceBERT⁶ from English to Danish. To this end, we used the Opus corpus⁷ English-Danish parallel corpus. English translations of the eDok documents were added to the corpus to augment its medical content. Given the input length limitations of most language models [4], and to avoid missing important semantic information from documents, embeddings were separately computed for each document section.

Query pre-processing. Spelling errors and incorrect information (for example, confusing two medication names that look or sound alike) might not only lead to retrieving irrelevant documents, but also to critical errors that could pose a risk to patients' well-being [9]. An inspection of historical queries showed that approximately 2% included misspelled words or Latin transliteration of characters. Given the low prevalence of spelling mistakes, we opted for using Hunspell⁸, one of the simplest and most popular correction

tools. The standard Danish dictionary was extended to include medical terms, classifications, and synonyms developed by the Danish Health Data Authority⁹ for documenting clinical and administrative conditions, English medical acronyms¹⁰ [5], and the eDok documents' vocabulary.

Query ensemble. Query recommendations are an integral part of retrieval systems to help users explore the information space by refining the current query, switching to a different search task, or diversifying the retrieved documents [11]. Here, we are interested in recommending alternative queries after users submit their query instead of query completion suggestions. These recommendations can assist users in clarifying their search, refining the query to be more accurate and concise, or making modifications based on examining the search results [12].

We had access to the eDok searched queries and the corresponding clicked documents for the period October 2022–February 2023. Finding related queries solely based on co-click (collaborative) behaviours was discarded due to the sparse nature of click history. Inspired by query2vec [7], which learns simultaneously from sparse clicks and the clicked documents to improve query embedding, we trained a FastText¹¹ model considering the combination of query and the title of the corresponding clicked document. FastText was chosen as it can naturally handle unknown terms, which, in this case, allows to analyze either unknown or variations of known queries. Then, given a query in search time, the top-k similar queries are retrieved. Similarities are computed using Faiss¹².

The widespread adoption of electronic health records has created a significant amount of diverse patient-related information (e.g., demographics, family history, past medical problems, medications, and current problems) in the form of free text that could be exploited for query refinement. These notes could be long, which, if queried as a whole, could be problematic, as the retrieval model will try to retrieve the documents that best adjust to the full content instead of what is truly important [20]. As a result, noisy terms could affect the overall retrieval performance. Several studies [3, 13] reported improvements in retrieval effectiveness when adding limited contextual concepts from clinical notes to queries. Similarly, preliminary analyses based on the historical eDok queries showed that extending the initial one-word query caused the clicked document to be ranked, on average, ten positions earlier. As there is yet no publicly available model for medical entity detection in Danish, concept extraction was based on noun phrase detection using spaCy. Then, we kept the words appearing in the eDok collection and recommended them to the user according to their frequency. Users will see a highlighted version of the input text, where concepts can be clicked and added to the query.

Document retrieval and recommendation. According to several studies (e.g., [21]), users tend to start searches using short queries (1.7 words on average), and then are likely to reformulate them into longer ones. This implies that the system is expected to receive queries of varying lengths and levels of contextualization, ranging from short and uncontextualized to long and meaningful

³The demo and the implementation can be accessed at: <https://linktr.ee/re2dan>.

⁴<https://pyterrier.readthedocs.io/en/latest/>

⁵<https://spacy.io/models/da>

⁶<https://www.sbert.net/>

⁷The corpus is available at: <https://opus.nlpl.eu/>. In particular, we considered: TED2020, WikiMatrix, ELRC-848-laegemiddelstyrelsen, ELRC-EMEA, ELRC-2877-EU_publications_medi

⁸<https://github.com/dvwright/phunspell>

⁹<https://sundhedsdata.item.dk/>

¹⁰<https://github.com/lisavirginia/clinical-abbreviations>

¹¹<https://github.com/facebookresearch/fastText>

¹²<https://github.com/facebookresearch/faiss>

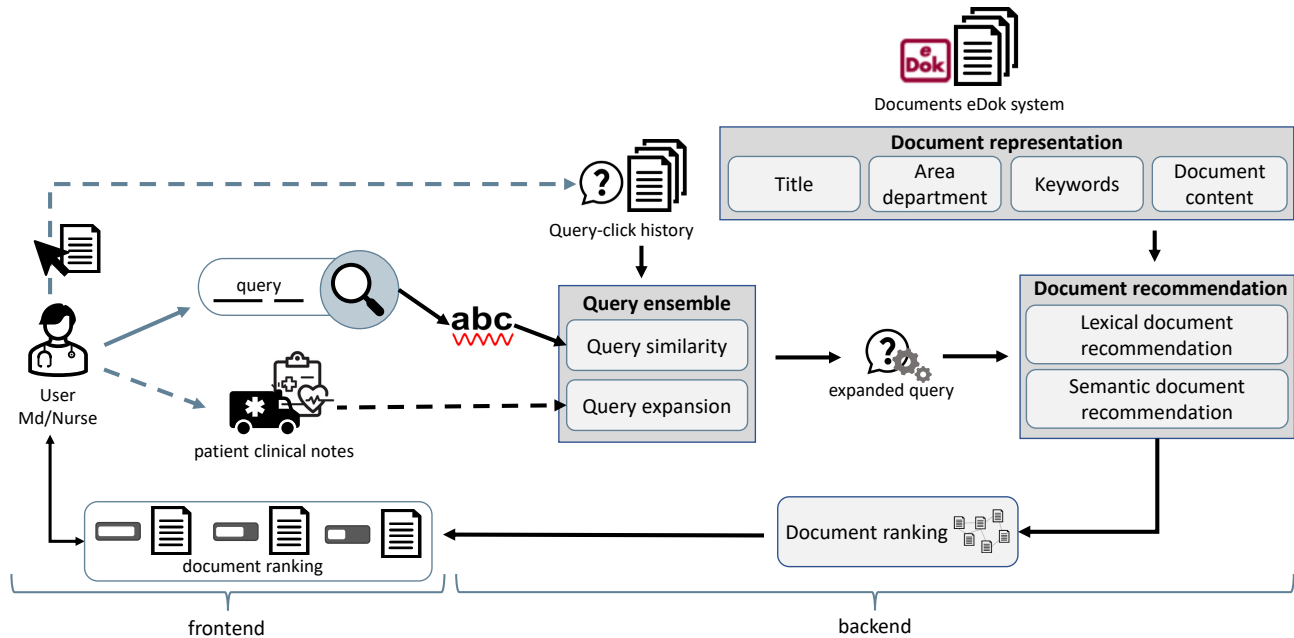


Figure 1: Re²Dan pipeline overview (Dashed lines indicate optional inputs)

queries. To handle this, two distinct retrieval models were implemented. Firstly, a lexical retrieval model was employed for single-word queries, based on the built inverted index. While this model is primarily designed for queries consisting of a single word, it also supports "and", "or" and phrase queries. Documents are retrieved based on their BM25 score [14]. Different weights could be assigned depending on whether the query term appears in the title or the content of the document. Second, a model based on the semantic similarity between phrase queries and document embeddings. To overcome limitations imposed by document length while capturing documents' context more effectively, similarities were computed at the document sub-unit level (i.e., sections or paragraphs) and then aggregated to obtain a final cumulative similarity [4].

Document ranking. Once documents are retrieved, they are ranked according to their similarity with the query. For multi-word queries, documents are retrieved by both models and then rankings are combined using Borda count [10]. By using this semantic model, preliminary analyses showed that the position of the clicked documents was improved for multi-word queries when compared to the original eDok system. Finally, the ranking is presented to the user. Document clicks are registered to augment the query-click history and improve similar query recommendations. Here, we consider clicks as an implicit feedback indicator of document relevance.

3 IMPLEMENTATION AND USER INTERFACE

The backend was implemented in Python as a Flask API¹³. Retrieved results and user clicks are stored in a MySQL¹⁴ database. For now, no personal information is stored from users, except for an anonymous

session id (which changes with every page full reload). The frontend was implemented using Streamlit¹⁵.

4 CONCLUSION AND FUTURE WORK

We have presented Re²Dan, a Danish medical document retrieval and recommender system. Re²Dan combines lexical and semantic searches, as well as recommendations of similar queries and medical concept extraction from related contextual information. Despite being publicly available, Re²Dan is still a beta prototype release. We plan to expand its functionality in the future by introducing personalization of document retrieval by considering the place of work and medical specialty of users, modelling medical entity recognition, and introducing personalization of contextual concept suggestion based on previous user choices (for example, suggesting only certain types of entities). Finally, we plan a quantitative evaluation of the system and a comprehensive user study.

ACKNOWLEDGMENTS

We gratefully acknowledge funding of this research by the Innovation Fund Denmark (IFD) through the Grand Solution project *Hospital@Night*.

REFERENCES

- [1] Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? *Advances in neural information processing systems* 27 (2014).
- [2] Jón Ingi Bergsteinsson. 2013. A checklist-based guideline system. *BMC Medical Informatics and Decision Making* 8, 1 (2013), 108.
- [3] Lijuan Diao, Hong Yan, Fuxue Li, Shoujun Song, Guohua Lei, and Feng Wang. 2018. The research of query expansion based on medical terms reweighting in medical information retrieval. *EURASIP Journal on Wireless Communications and Networking* 2018 (2018), 1–7.
- [4] Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. 2021. Self-Supervised Document Similarity Ranking via Contextualized Language

¹³<https://flask.palletsprojects.com>

¹⁴<https://www.mysql.com>

¹⁵<https://streamlit.io>

- Models and Hierarchical Inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 3088–3098. <https://doi.org/10.18653/v1/2021.findings-acl.272>
- [5] Lisa Grossman Liu, Raymond H Grossman, Elliot G Mitchell, Chunhua Weng, Karthik Natarajan, George Hripcsak, and David K Vawdrey. 2021. A deep database of medical abbreviations and acronyms for natural language processing. *Scientific Data* 8, 1 (2021), 149.
 - [6] William Hersch. 2008. *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media, New York, NY, USA.
 - [7] Dongyeop Kang and Inho Kang. 2015. *Query2vec: Learning deep intentions from heterogeneous search logs*. Technical Report. Technical report, Carnegie Mellon University.
 - [8] Bevan Koopman, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2011. Evaluating Medical Information Retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) (SIGIR '11). Association for Computing Machinery, New York, NY, USA, 1139–1140. <https://doi.org/10.1145/2009916.2010088>
 - [9] Kenneth H Lai, Maxim Topaz, Foster R Goss, and Li Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of biomedical informatics* 55 (2015), 188–195.
 - [10] Shili Lin. 2010. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 5 (2010), 555–570.
 - [11] Agnès Mustar, Sylvain Lamprier, and Benjamin Piwowarski. 2021. On the study of transformers for query suggestion. *ACM Transactions on Information Systems (TOIS)* 40, 1 (2021), 1–27.
 - [12] Umut Ozertem, Olivier Chapelle, Pinar Donmez, and Emre Velipasaoglu. 2012. Learning to Suggest: A Machine Learning Framework for Ranking Query Suggestions. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) (SIGIR '12). Association for Computing Machinery, New York, NY, USA, 25–34. <https://doi.org/10.1145/2348283.2348290>
 - [13] Georgios Peikos, Daria Alexander, Gabriella Pasi, and Arjen P. de Vries. 2023. Investigating the Impact of Query Representation on Medical Information Retrieval. In *Advances in Information Retrieval*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer Nature Switzerland, Cham, 512–521.
 - [14] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.
 - [15] Salimah Z Shariff, Shayna AD Bejaimal, Jessica M Sontrop, Arthur V Iansavichus, R Brian Haynes, Matthew A Weir, and Amit X Garg. 2013. Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches. *Journal of medical Internet research* 15, 8 (2013), e2624.
 - [16] Romina Sharifpour, Mingfang Wu, and Xiuzhen Zhang. 2023. Large-scale analysis of query logs to profile users for dataset search. *Journal of Documentation* 79, 1 (2023), 66–85.
 - [17] Lynda Tamine and Cecile Chouquet. 2017. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing & Management* 53, 2 (2017), 332–350.
 - [18] Lynda Tamine, Cécile Chouquet, and Thomas Palmer. 2015. Analysis of biomedical and health queries: Lessons learned from TREC and CLEF evaluation benchmarks. *Journal of the Association for Information Science and Technology* 66, 12 (2015), 2626–2642.
 - [19] Lynda Tamine and Lorraine Goeuriot. 2021. Semantic information retrieval on medical texts: research challenges, survey, and open issues. *ACM Computing Surveys (CSUR)* 54, 7 (2021), 1–38.
 - [20] Yue Wang and Hui Fang. 2018. Key Terms Guided Expansion for Verbose Queries in Medical Domain. In *Information Retrieval Technology*, Yuen-Hsien Tseng, Tet-suya Sakai, Jing Jiang, Lun-Wei Ku, Dae Hoon Park, Jui-Feng Yeh, Liang-Chih Yu, Lung-Hao Lee, and Zhi-Hong Chen (Eds.). Springer International Publishing, Cham, 143–156.
 - [21] Lei Yang, Qiaozhu Mei, Kai Zheng, and David A Hanauer. 2011. Query log analysis of an electronic health record search engine. In *AMIA annual symposium proceedings*, Vol. 2011. American Medical Informatics Association, Bethesda, Maryland, USA, 915.