Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval

Priyanka Mandikal, Raymond Mooney

The University of Texas at Austin, USA

Abstract

Traditional information retrieval is based on sparse bag-of-words vector representations of documents and queries. More recent deep-learning approaches have used dense embeddings learned using a transformer-based large language model. We show that on a classic benchmark on scientific document retrieval in the medical domain of cystic fibrosis, that both of these models perform roughly equivalently. Notably, dense vectors from the state-of-the-art SPECTER2 model do not significantly enhance performance. However, a hybrid model that we propose combining these methods yields significantly better results, underscoring the merits of integrating classical and contemporary deep learning techniques in information retrieval in the domain of specialized scientific documents.

Keywords

scientific document retrieval, information retrieval, hybrid methods

1. Introduction

The field of information retrieval (IR) has undergone significant transformations over the years, evolving from traditional methods based on sparse vector representations to more advanced techniques utilizing deep learning models. Historically, the vector-space model in IR, as described by Manning et al. [1], has been predominant. This model represents documents and queries using sparse, Term-Frequency/Inverse-Document-Frequency (TF/IDF) weighted bag-of-words (BOW) vectors, a method known for its simplicity and effectiveness in various text retrieval applications.

With the advent of deep learning, there has been a paradigm shift towards utilizing dense embeddings, especially those derived from transformer-based large language models (LLMs). These methods, exemplified by Dense Passage Retrieval (DPR) [2], offer a more nuanced and context-aware approach to text representation, often resulting in improved retrieval performance compared to traditional sparse-vector methods.

In the specialized field of scientific document retrieval, dense-vector approaches have gained prominence. One notable example is the SPECTER model [3], which uses a BERT-based framework for embedding scientific texts. It employs contrastive learning, leveraging citation information for supervision, to identify document similarities effectively. On a variety of tasks including document classification, citation prediction, and recommendation, SPECTER has been shown to

The 4th Workshop on Scientific Document Understanding, February 26, 2024, Vancouver, BC, Canada
mandikal@utexas.edu (P. Mandikal); mooney@utexas.edu (R. Mooney)
https://priyankamandikal.github.io/ (P. Mandikal); https://www.cs.utexas.edu/~mooney/ (R. Mooney)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

outperform a variety of competing methods [3]. The model's recent iteration, SPECTER2 [4], further refines this approach by training on an expanded corpus of 6M triplets spanning a wider array of scientific fields, producing more representative document embeddings.

In this paper, we evaluate SPECTER2 on a classic benchmark in scientific document retrieval in the medical area of cystic fibrosis [5]. Contrary to the anticipated superiority of dense embeddings, our findings reveal that the performance of SPECTER2 is comparable, and in some cases, marginally inferior to that of traditional sparse-vector retrieval methods. This observation led us to explore a hybrid retrieval model that combines the strengths of both sparse and dense vector representations. This simple, yet elegant hybrid approach performs remarkably well, clearly outperforming the base models on both standard precision/recall and NDCG metrics. Through our research, we aim to determine the optimal blend of these approaches and demonstrate the potential benefits of such a hybrid model in enhancing document retrieval accuracy and efficiency.

Our work contributes to the evolving landscape of IR by not only comparing traditional and modern IR techniques but also by demonstrating the efficacy of a hybrid approach in a specialized medical domain. This study not only contributes to the existing body of knowledge in IR but also provides practical insights for the development of more effective retrieval systems, particularly in specialized fields such as medical research. The findings have implications for both the theoretical understanding of IR models and their application in real-world scenarios.

2. Related Work

Classic IR. The evolution of information retrieval (IR) techniques, particularly in the context of scientific document retrieval, has seen a significant shift from traditional sparse vector models to sophisticated deep learning approaches. The seminal work of Salton and McGill [6] laid the foundation for vector-space models, employing bag-of-words (BOW) representations and TF/IDF weighting, a standard in early IR systems. This approach has been extensively studied and optimized over the years, as documented by Manning et al. [1].

Deep learning based IR. The advent of deep learning has introduced a new paradigm in IR, with dense vector embeddings offering nuanced text representations. Pioneering this shift, Le and Mikolov [7] demonstrated the effectiveness of dense embeddings with their work on Doc2Vec, showing significant improvements in capturing semantic text similarities.

Large Language Models. The development of transformer-based models, as introduced by Vaswani et al. [8], further revolutionized this field, enabling more complex and contextually rich text representations. Focusing on scientific documentation, the SPECTER model [3], based on BERT architecture [9], emerged as a significant advancement, utilizing citation networks for improved document embeddings. Its successor, SPECTER2 [4], expanded on this with a larger training corpus and enhanced learning strategies. Recent research has also explored the integration of sparse and dense vector methods [10, 11]. However, such a hybrid approach has not been previously explored for scientific document retrieval. Such hybrid models are particularly relevant in scientific domains where traditional IR techniques remain competitive, as observed in our study focusing on cystic fibrosis document retrieval.

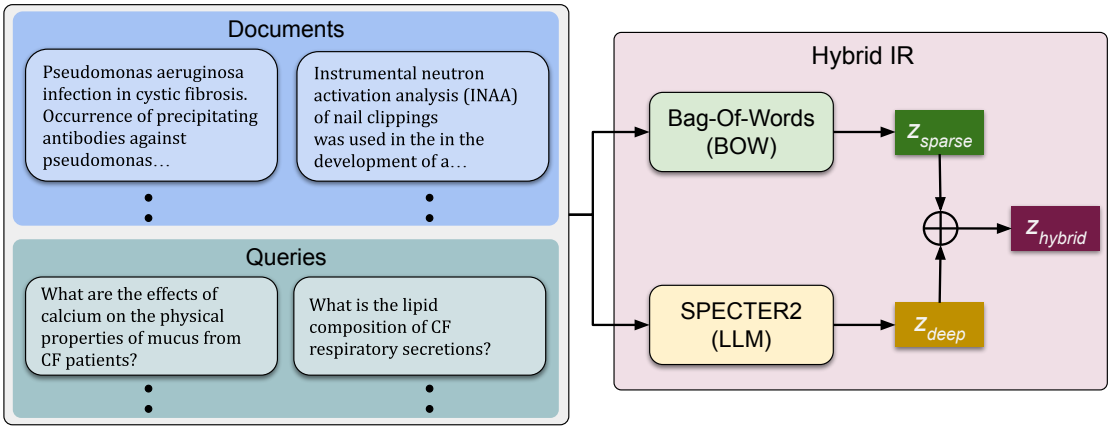

Figure 1: Overview of our approach. On a medical dataset of cystic fibrosis documents, we combine sparse bag-of-words embeddings with dense embeddings from a SOTA LLM (Specter2 [4]) to produce a hybrid retriever that significantly outperforms both methods.

3. Dataset

The Cystic Fibrosis Database (CF) [5] consists of 1,239 documents published from 1974 to 1979 discussing Cystic Fibrosis, and a set of 100 queries with the respective relevant documents as answers. It is a classic IR benchmark that is discussed in an early textbook [12]. The documents are represented by an abstract and lists of minor and major subjects, but do not include the full text of the document. The queries are full sentence questions, e.g. "Can one distinguish between the effects of mucus hypersecretion and infection on the submucosal glands of the respiratory tract in CF?" Each query has been annotated with ratings from four domain experts, where each document is assigned a relevance score: 0: not relevant, 1: marginally relevant, 2: highly relevant.

When measuring precision and recall, any document rated as at least marginally relevant by at least one annotator is considered relevant. To provide a more nuanced gold-standard, the average of the judges' ratings is also used as a continuous relevance value for each document. This is used when measuring Normalized Discounted Cumulative Gain (NDCG), a standard IR metric that utilizes continuous gold-standard relevance ratings [13, 1].

4. Approach

We present an overview of our hybrid retriever combining sparse and deep embeddings in Fig. 1 and describe each component below.

4.1. Sparse Retrieval Model

To generate sparse embeddings, we use a simple, classical, bag-of-words vector-space retrieval (VSR) method as a baseline. It uses Term-Frequency/Inverse-Document-Frequency (TF/IDF)

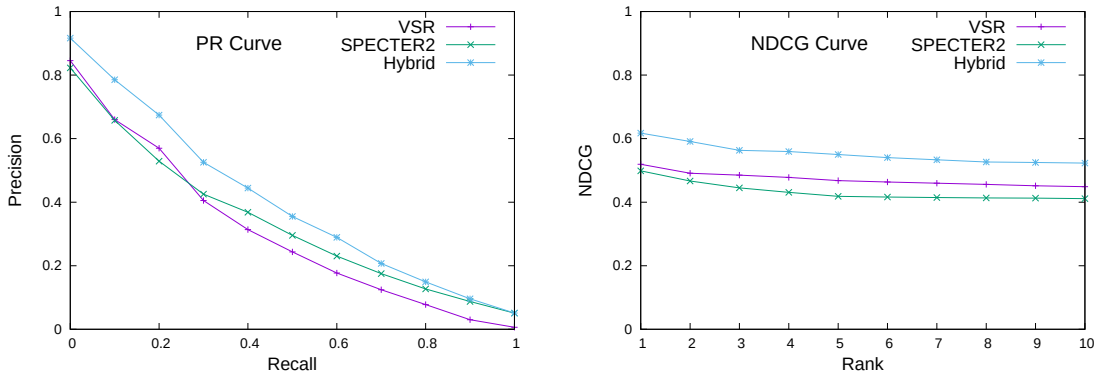

Figure 2: Results on the Cystic-Fibrosis dataset. The hybrid approach ($\lambda = 0.8$) outperforms both traditional sparse vector-space retrieval (VSR) and state-of-the-art deep embeddings (SPECTER2 [4]) in both PR (left) as well as NDCG (right) metrics.

token weighting and cosine similarity to measure query/document relevance [12, 1]. A traditional inverted index is used to efficiently retrieve documents that share tokens with the query after removing a standard set of stopwords.

4.2. Dense Retrieval Model

To generate deep embeddings, we use SPECTER2 [4] as a representative of a state-of-the-art deep-learning model explicitly trained to produce effective dense-vector representations of scientific documents. SPECTER2 is a successor to SPECTER [3] which in turn was a followup to SciBERT [14]. It uses a transformer-based LLM that was explicitly trained on a large corpus of scientific documents from the Semantic Scholar search engine [15].

SPECTER2 uses additional self-supervised training based on citations to produce good document abstract embeddings. A contrastive learning approach is used to encourage similar 768-dimensional embeddings for an article its direct citations. SPECTER2 has been trained on 6M triplets spanning 23 fields of study. Our retrieval model ranks documents based on a similarity score between the dense vectors generated for a query and the document abstract.

4.3. Hybrid Retrieval Model

Finally, we combine the sparse and dense retrieval models to produce a hybrid retriever (see Fig. 1). Our hybrid model uses a simple weighted combination of the query/document similarities in the sparse and dense embedding spaces. Specifically, we score a document D for query Q using the ranking function:

$$\begin{aligned} S_{hybrid}(D,Q) = &\lambda\ Sim\big(z_{dense}(D), z_{dense}(Q)\big) + \\ &(1-\lambda)\ Sim\big(z_{sparse}(D), z_{sparse}(Q)\big) \end{aligned} \tag{1}$$

Where z_{dense} and z_{sparse} denotes the dense and sparse embedding functions and Sim is cosine similarity measuring the angle between such vector embedddings. In our experiments we present results for different values of the hyperparameter λ.

Query: How may heterozygotes for CF be identified?

Method	P@1	P@3	P@5	P@10
VSR	**100**	33.33	60	60
SPECTER2	**100**	66.67	60	50
Hybrid	**100**	**100**	**80**	**70**

Sample Retrieval

…sodium in nail clippings for diagnosing cystic fibrosis (CF) in children and adults, for detecting heterozygotes and for screening in the neonatal period. Nail clippings from 1322 newborns, 22 CF patients (two of them newborns), 52 healthy controls and 22 heterozygotes were analyzed. The discrimination between CF patients and controls was found to be precise for individuals above one year of age and INAA of nail clippings should be accepted as a diagnostic test for CF after this age. Heterozygotes could not be detected…

Query: How effective is bronchial lavage in CF patients?

Method	P@1	P@3	P@5	P@10
VSR	**100**	66.67	60	50
SPECTER2	**100**	66.67	60	40
Hybrid	**100**	**100**	**80**	**70**

Sample Retrieval

In an attempt to retard deterioration of pulmonary function, a young man with cystic fibrosis underwent broncho-pulmonary lavage of an entire lung. In spite of theoretical advantages, the lavage produced an acute deterioration in the patient's clinical and physiologic status, from which he had not recovered one month later. We conclude that bronchopulmonary lavage in cystic fibrosis may be hazardous and should be reserved for those cases in which conventional methods of improving tracheobronchial clearance have failed.

Query: Do fibroblasts from CF patients grow at a normal rate?

Method	P@1	P@3	P@5	P@10
VSR	0	0	20	20
SPECTER2	**100**	33.33	40	50
Hybrid	**100**	**66.67**	**80**	**60**

Sample Retrieval

This paper analyzes the cell cycle of cystic fibrosis-derived fibroblasts and statistically demonstrates their failure to support a normal population doubling time in vitro and indicates that this is due largely to a failure of the cystic fibrosis-derived cells to enter the cell cycle in normal numbers. We have no explanation for the significant difference in the behavior of these cells from patients with cystic fibrosis vis-a-vis either normal or disease control fibroblasts, save to conclude that it proceeds from an intrinsic property of the cells themselves.

Query: What is the prognosis for patients who have had an episode of respiratory failure?

Method	P@1	P@3	P@5	P@10
VSR	0	33.33	**40**	40
SPECTER2	0	0	20	20
Hybrid	**100**	**66.67**	**40**	**60**

Sample Retrieval

The diagnosis, management and prognosis of 17 infants with cystic fibrosis (CF) and severe respiratory disease were reviewed for the period 1968 to 1972. The clinical course of these infants was characterized by a bronchiolitis-like syndrome with failure to thrive and malnutrition… Despite vigorous therapy including antibiotics there was a 60% mortality. A delay in the diagnosis of CF from the onset of respiratory symptoms with a mean of six weeks was considered an important factor affecting survival. This data supports the need for developing a reliable screen test for CF at birth.

Figure 3: Sample queries and retrievals. We show results for the three methods on four sample queries. Our hybrid approach outperforms both VSR and SPECTER2, retrieving more number of relevant documents higher in the retrieval ranks. We also show sample retrievals relevant to the given query with keywords highlighted.

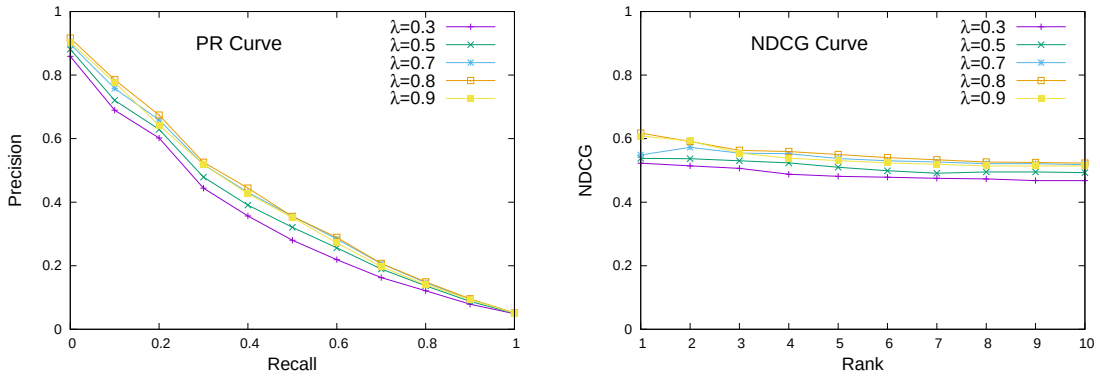

Figure 4: Ablations with different values of λ. Weighing the deep embedding with a weight of $\lambda = 0.8$ (where sparse gets a weight of 0.2) produces the best results for both PR and NDCG metrics.

5. Experiments

In this section, we provide empirical results for the effectiveness of hybrid embedding and also perform ablations on various hyperparameter and model variants.

Effectiveness of hybrid embeddings. We compare the sparse, dense and hybrid retrieval models on the CF corpus using standard IR experimental methods [1]. We plot precision-recall (PR) curves using a standard approach that interpolates retrieval precision for a set of standard recall levels [12]. We also compute NDCG values [13] for the top 10 retrievals using the continuous gold-standard relevance ratings determined from the average human evaluator ratings. We use cosine similarity as our similarity metrics while comparing the document and query embeddings. We also tried using Euclidean distance to compare deep embeddings, but it performed almost identically to cosine similarity.

Quantitative results are shown in Fig. 2. We observe that the hybrid approach ($\lambda = 0.8$) outperforms both traditional sparse embeddings (VSR) and state-of-the-art deep embeddings (SPECTER2 [4]) in both PR as well as NDCG metrics. We also show results for four sample queries in Fig. 3. We can see that while both VSR and SPECTER2 fail to consistently retain high precision values for their retrievals, our hybrid approach is able to effectively retrieve greater number of highly relevant documents higher up in the retrieval list. We also depict sample retrievals with keywords highlighted for the given queries to illustrate the relevance of the retrieved documents to the user query. This underscores the simple yet effective strategy of combining sparse and dense embeddings for scientific document retrieval.

Hyperparameter λ. We conduct ablations on the value of λ for effective combination of the dense and sparse embeddings. Results are depicted in Fig. 4. We observe that weighing the deep embedding with a weight of $\lambda = 0.8$ (where sparse gets a weight of 0.2) produces the best results for both PR and NDCG metrics. This indicates that a higher emphasis on the dense embeddings from the SPECTER2 model, while still incorporating a significant contribution from the traditional sparse BOW vectors, achieves an optimal balance for document retrieval

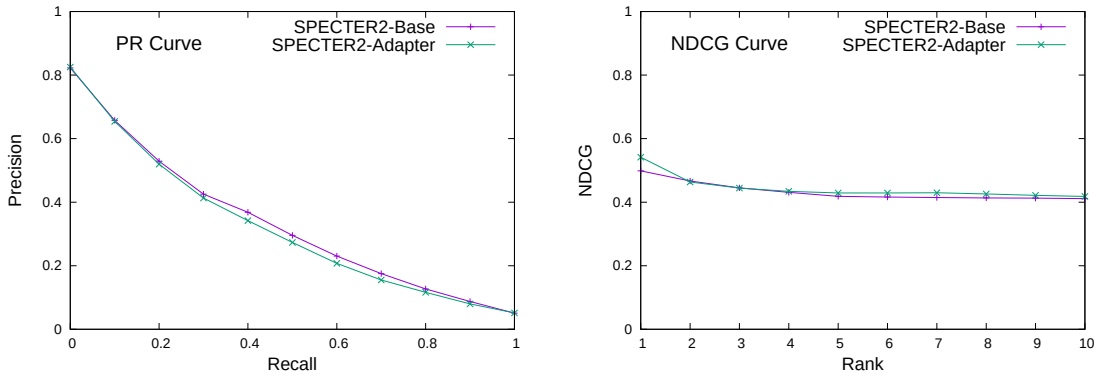

Figure 5: Ablations with SPECTER2 base vs adapter versions. The base model performs as well as the adapter variant on the PR metric. The adapters marginally improve performance on NDCG but hurt precision for high recall levels .

in the cystic fibrosis dataset. This hybrid approach effectively captures the strengths of both methodologies, utilizing the contextual richness and semantic understanding from the dense embeddings, and the keyword-focused precision of sparse vectors. Therefore, our findings underscore the importance of a nuanced combination of these models, tailored to the specific characteristics of the dataset and the retrieval task at hand. While scaling this hybrid approach to different datasets, this hyperparameter can be set using normal grid-search on a validation set.

Base vs Adapters. SPECTER2 includes adapters that are explicitly trained to produce improved representations of abstracts and queries, but we found that they did not improve performance over the base model on the cystic-fibrosis dataset. This ablation is shown in Fig. 5. The base model performs as well as the adapter variant on the PR metric. The adapters only marginally improve performance on NDCG while marginally decreasing precision at higher recall levels. Therefore, our results in Figs. 2 and 4 utilize the SPECTER2 base model.

6. Conclusion

This paper has explored applying SOTA deep embeddings for scientific documents using the SPECTER2 model to a classic benchmark dataset in the domain of medical document retrieval. We found that they gave disappointing results, at best performing only about as well as traditional sparse TF/IDF vector-space retrieval. However, a simple hybrid approach that combines sparse and dense retrieval gave significantly improved results. The findings align with the growing consensus in the literature that the future of IR lies in the thoughtful integration of classical and modern methodologies.

References

[1] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, 2008.

[2] V. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, W. Yih, Dense passage retrieval for open-domain question answering, EMNLP (2020). URL: https://arxiv.org/abs/2004.04906. arXiv:2004.04906.

[3] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, SPECTER: document-level representation learning using citation-informed transformers, ACL (2020). arXiv:2004.07180.

[4] A. Singh, M. D'Arcy, A. Cohan, D. Downey, S. Feldman, SciRepEval: A multi-format benchmark for scientific document representations, 2023. arXiv:2211.13308.

[5] W. Shaw, J. Wood, R. Wood, H. Tibbo, The cystic fibrosis database: Content and research opportunities, Library & Information Science Research 13 (1991).

[6] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., 1983.

[7] Q. Le, T. Mikolov, Distributed representations of sentences and documents, Proceedings of the 31st International Conference on Machine Learning (ICML) (2014).

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems (2017).

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019).

[10] M. G. Arivazhagan, L. Liu, P. Qi, X. Chen, W. Y. Wang, Z. Huang, Hybrid hierarchical retrieval for open-domain question answering, in: Findings of the Association for Computational Linguistics: ACL, 2023.

[11] Y. Luan, J. Eisenstein, K. Toutanova, M. Collins, Sparse, Dense, and Attentional Representations for Text Retrieval, Transactions of the Association for Computational Linguistics (TACL) (2021).

[12] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, New York, 1999.

[13] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems 20 (2002).

[14] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, EMNLP (2019). arXiv:1903.10676.

[15] AI2, Semantic scholar, www.semanticscholar.org (2015).