Searching for Fine-Grained Queries in Radiology Reports Using Similarity-Preserving Contrastive Embedding

Tanveer Syeda-Mahmood STF@US.IBM.COM
IBM Almaden Research Center
650 Harry Road
San Jose, CA, USA

Luyao Shi LUYAO.SHI@IBM.COM
IBM Almaden Research Center
650 Harry Road
San Jose, CA, USA

Editor: Editor's name

Abstract

The ability to search in unstructured reports of electronic health records requires tools that can recognize clinically meaningful fine-grained descriptions both in queries and in report sentences. Existing methods of searching reports that use either information retrieval or deep learning techniques to model use context, lack an inherent understanding of the clinical concepts or their variants that capture the same underlying clinical semantics. In this paper, we present a new search algorithm that combines principles of information retrieval and deep learning-driven textual encoding approaches with natural language analysis of sentences in reports for fine-grained descriptors of concepts. In particular, we learn a clinical similarity-preserving embedding from a chest X-ray lexicon using a new contrastive loss. This allows us to form a report index that is robust to different forms of expressing for clinical concepts in queries. The results show marked improvement in the quality of retrieved reports as judged through average recall and mean average precision over a broad range of difficult queries.

1. Introduction

With the wide-scale adoption of electronic health records (EHR) as the main repositories to house structured and unstructured data in hospitals, tools for performing clinically relevant searches have become important for clinicians, staff, and researchers alike. Such tools could aid in a variety of use cases such as clinical decision support where patients with similar conditions are searched (Syeda-Mahmood, 2010), auditing and review to see compliance with care practices (Guo et al., 2018), or cohort selection for patients satisfying inclusion and exclusion clinical criteria of clinical trials (Spasic et al., 2019). Other secondary use purposes such as quality assurance, population health management, and clinical and translational research have also been found made possible due to search capabilities in EHR systems (Hanauer et al., 2015). Finally, searching in reports has also become important recently for auto-labeling of image datasets from their companion reports for building deep learning models (Syeda-Mahmood et al., 2020).

Despite the importance of this problem, it remains challenging for several reasons. First, the terms used in the query may be ad hoc and no direct matches may be present in the reports. Secondly, spelling variants and other synonymous forms may be present. When the query is fine-grained, i.e. asks not only for the finding but also specifies the laterality, location, etc. (e.g. "right lower lobe pneumonia") then the search method should be able to identify key clinical concept in both queries and report sentences(e.g. "pneumonia"). Further, it should be able to associate the relevant modifiers with the core finding (e.g. "right lower lobe"). Finally, the search methods should be relatively fast limiting the number of query expansions that can be done live during search.

Popular methods address this problem using mainly document-centric (Robertson and Zaragoza, 2009) or neural vector-based (Reimers and Gurevych, 2019) approaches. In the document-centric approach, the report text is broken into sentences and their words tokenized. In the neural vector approach, the entire sentence or its fragment is encoded as a neural vector. In the former method, search is achieved through document ranking methods using the ratio of term frequency to inverse document frequency (TF/IDF). In the neural vector approach, search is achieved through cosine distance in vector space. Neither approach has a good understanding of the clinical semantics in terms of focusing on the core clinical concept and its fine-grained description, their natural spoken variations, nor equivalent terms that mean the same as the clinical concept occurring possibly in different use context.

To illustrate this, let us consider the type of responses returned by these methods to search queries on a database of 2,770 chest X-ray reports drawn from the Indiana University's collection (Demner-Fushman et al., 2016). For a query such as "right lower lobe pneumonia" shown in Table 1, the matching sentences from the corresponding reports are shown in Table 1 Column 1 and 2 respectively for the two methods. As can be seen from the ranked list produced, there are several mismatches indicating a basic lack of understanding of the clinical context where the meaning of the core finding or the association of relevant modifiers with core finding is lost. For example, there are negated instances of pneumonia (e.g. "no active pneumonia") or partial matches to terms such as "right upper lobe" which is not the main focus of the query. The neural IR approach (Reimers and Gurevych, 2019) also returned results that has little resemblance to the query.

Table 2 shows another example where term similarity in meaning was expected to be observed. Here the query is "fluid overload" and the intention was to capture cases of pulmonary edema or vascular congestion. While the TF/IDF's top match is one of the matching sentences, it is purely a coincidence due to the match in the term "overload" rather than due to any semantic understanding of the condition. The neural IR matches are also inconsistent at best, lacking an understanding of what was significant to capture in the query from a match perspective. Furthermore, since embedding methods average the vectors from each of the words to form phrasal vectors, they may not necessarily match those derived from the shorter query phrases.

Thus we see that to build a robust search engine for clinical documents, there should be a strong understanding of the clinical concepts including their synonymous ways of expressing, a good understanding of negations, as well as an emphasis on fine-grained descriptors. In this paper we present an approach that addresses many of these deficiencies of existing search techniques. Specifically we adopt a hybrid approach where we combine principles of

Query="right upper lobe pneumonia" using TF/IDF	**Query="right upper lobe pneumonia" using ClinicalBERT**
1. No active penumonia.	1. Left-sided cardiomediastinal contours are obscured by collapse of the left lung.
2. Left upper lobe calcified granuloma noted.	2. Subsegmental atelectasis verus scarring in the right midlung and left lower lobe.
3. Stable calcified granuloma in the right upper lobe.	3. Subsegmental atelectasis in the left lower lung.
4. Interval development of the mild patchy airspace opacities within the posterior aspect of the right upper lobe, concerning for underlying penumonia.	4. Granulomatous mediastinal calcifications.
5. No change in the small calcified granuloma in the right upper lobe.	5. Stable atelectatic/fibrotic changes of the visualized lung, and stable left-sided calcified granuloma.
6. There is stable appearing left upper and right upper lobe bullous disease.	6. Bronchovascular crowding without typical findings of pulmonary edema.
7. Stable right upper lobe calcified granuloma.	7. Bronchovascular crowding, indistinct central vascular margination.
8. In the collapsed left upper lobe are stranding and penumatoceles.	8. Bibasilar pleural scarring.
9. There is round density within the anterior segment of the right upper lobe.	9. Right-sided chest xxxx catheter tip is at the lower svc.
10. Anterior segment of upper lobe, rounded focal density.	10. Left-sided tunneled catheter terminates at the caval atrial junction.

Table 1: Illustration of the difficulty of searching for clinical concepts in radiology reports. Methods fail to recognize what the key clinical concept is for which a match should be found. They also do not pair it with relevant modifiers.

information retrieval and deep learning-driven textual encoding approaches with the natural language analysis of sentences in reports for fine-grained descriptors of concepts. Specifically, we capture the semantic context by combining vocabulary-driven concept extraction with natural language analysis of the sentence structure to extract key clinical concepts along with their associated description modifiers to form fine-grained finding (FFL) patterns that include negations. The method assumes a domain lexicon or ontology that captures similar meaning terms found in the FFL patterns. All such pairs of terms are used to learn a contrastive neural embedding such that terms closer in meaning and sense are projected close together. All the constituent terms of the FFL patterns from the report sentences are used to form an index linking the terms back to their FFL patterns, their enclosing sentences and hence their enclosing report documents.

Given a new query representing a phrase, a similar sentence analysis is performed to extract the FFL pattern from the query. Each element of the query FFL pattern is then projected into the contrastively learned embedding to retrieve the nearest in meaning term

Query="fluid overload" using TF/IDF	**Query="fluid overload" using ClinicalBERT**
1. Increased interstitial lung markings are seen, possibly due to volume overload.	1. Hyperinflation lungs.
2. No significant change in pneumothorax or right pleural fluid.	2. Hyperinflation is present.
3. No pleural fluid collection or pneumothorax.	3. Hyperinflation of the lungs.
4. No pleural fluid or pneumothorax is appreciated.	4. Stable hperinflation without focal alveolar consolidation.
5. No pneumothorax or visible pleural fluid.	5. Stable hyperinflation, bilateral upper lobe pleuroparenchymal near and nodular irregularities, right greater than left, xxxx opacities in the peripheral right lung most compatible with scarring.
6. No visible pleural fluid.	6. Bronchovascular crowding without typical findings of pulmonary edema.
	7. Bronchovascular crowding, indistinct central vascular margination.
	8. Emphysema and chronic changes are identified.
	9. Emphysema.

Table 2: Illustration of the difficulty of searching for clinical concepts in radiology reports. A second query case.

vector coming from any database reports. The results are aggregated across all query FFL pattern elements and ranked to yield overall matches to query. By using a meaning preserving embedding, we are able to match to similar terms yielding higher sensitivity, while the overall usage context is still captured in FFL pattern to yield higher specificity in matching patterns.

Generalizable Insights about Machine Learning in the Context of Healthcare

The use case analyzed in this paper indicates a good example of why incorporation of domain semantics is important to improve the performance of machine learning. Even though the current neural textual embedding approaches are indeed trained on clinical documents, they cannot identify the core clinical concept, its nearest meaning term, nor the association of fine-grained descriptions to their core clinical concept. The approach presented here that combines linguistic structure knowledge (i.e. sentence parsing) with meaning-preserving clinical domain semantics embedding within a neural or document retrieval framework shows how machine learning approaches can be augmented with knowledge for improved recognition. By swapping the domain lexicons for building the knowledge embedding, this method

could be applicable for search of other domain-specific document databases using neural IR approaches.

2. Related Work

A number of approaches have tried to address the problem of searching unstructured data in electronic health records (Natarajan et al., 2010; Meystre et al., 2008), and implementing them in hospitals (Hanauer et al., 2015). While the early engines attempted to do full text indexing and search, later versions offered refined search modeling inclusion and exclusion criteria (Hanauer et al., 2015). There has also been considerable work emphasizing context modeling such as negations in radiology reports for critical findings (Lacson et al., 2012). Recently, deep learning-based textual embedding methods have become available that are being applied to medical text. These methods aim to capture the use context of words in sentences of text as a surrogate for semantics or meaning (Bartusiak et al., 2019; Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2018). However, these methods cannot guarantee the preservation of meaning during retrieval especially when the query terms do not provide sufficient context as can happen in clinical text search. Nevertheless, deep learning-based NLP models have become popular for medical text being used for medical text classification, named entity recognition, medical question answering, de-identification of text or patient phenotyping (Spasic et al., 2020). More recently, these embeddings are trained on clinical documents to produce specialized versions such as clincalBERT (Huang et al., 2019) or BioBERT (Lee et al., 2020) which is a domain-specific language representation model pre-trained on large-scale biomedical corpora. As we showed in Table 1 and 2, while individual methods may address some aspect of the problem, such as negation, or word context, there is currently no approach that has a full understanding of clinical semantics, particularly, for searching radiology reports.

3. Methods

Our search algorithm has 4 main stages of processing, namely, (a) fine-grained concept extraction, (b) synonym expansion using supervised contrastive learning, (c) report index creation (d) search using contextual encoding. While our approach is generally applicable to any report collections that is covered by a clinical knowledge such as UMLS metathesaurus, our current implementation is illustrated in the context of chest X-ray radiology reports based on a chest X-ray ontology/lexicon recently reported in (Wu et al., 2020).

3.1. Fine-grained concept description in sentences

Following the approach described in (Syeda-Mahmood et al., 2020), we adopt the descriptor $F_i =< T_i|N_i|C_i|M_i^* >$ to describe any concept extracted from a sentence where F_i is the fine-grained label called FFL pattern, T_i is the finding type, $N_i = yes|no$ indicates a positive or ruled out finding, C_i is the core concept itself, and M_i are one or more of the possible finding modifiers. For example, to describe an anatomical finding of "left lower lobe pneumonia", we use the FFL pattern *disease|yes|pneumonia|left|lower lobe*. For chest radiology reports, the concept types are adequately covered by six major categories namely, anatomical findings, tubes and lines and their placements, external devices, viewpoint-related issues, and implied

Concept	Type	Anatomy	Synonymous phrases
Mediastinitis	Disease	Mediastinum	mediastinitis, mediastinitis/acute, mediastinum inflammations, inflammatory disorder of mediastinum, mediastinitides, inflammation of mediastinum
Shoulder disorder	Disease	Bones/Soft Tissues	sprengel deformity, shoulder/adhesive capsulitis, shoulder/frozen, shoulder joint disease, milwaukee shoulder/pseudogout syndrome, disorder of shoulder, shoulder disorders

Table 3: Illustration of the variations in the description of a finding across radiology reports.

diseases associated with findings. For the purposes of our experiments in this paper, we focus on disease concepts as those are the most commonly searched in electronic health record (Natarajan et al., 2010). The values taken by each of the above variables is derived from a chest x-ray lexicon as reported in an earlier work (Wu et al., 2020). Currently, the lexicon consists of over 11,000 unique terms covering 237 concepts that include anatomical findings, diseases, laterality, location, severity, and other appearance modifiers. Table 3 shows a few entry rows in the lexicon. As can be seen, each entry represents a core concept and lists potentially synonymous ways in which the concept could be described in reports. These are unordered lists curated by clinicians using a semi-supervised domain learning assistant (DLA) tool described in (Wu et al., 2020).

3.2. Extraction of FFL Labels from reports

To automatically extract such patterns from sentences, we use the overall concept extraction with phrasal grouping algorithm described in an earlier work (Syeda-Mahmood et al., 2020). Briefly, the algorithm for extracting FFL labels from sentences in reports consists of 4 steps, namely, (a) core finding and modifier detection, (b) phrasal grouping, (c) negation sense detection, (d) pattern completion. The vocabulary of core findings from lexicon and their synonyms was used to detect core concepts in sentences of reports using the vocabulary-driven concept extraction algorithm described in (Guo et al., 2017).

To associate modifiers with relevant core findings, we used a natural language parser called the ESG parser (McCord et al., 2012) which performed word tokenization and morpholexical analysis to create a dependency parse tree for the words in a sentence as shown in Figure 1. The initial grouping of words is supplied directly by the parse tree such as the grouping of terms “alveolar” and “consolidation” into one term “alveolar consolidation” shown in Figure 1. Further phrasal grouping is done by clustering the lemmas using word identifiers specified in the dependency tree. For this, a connected component algorithm is used on the word positions in slots, skipping over unknowns (marked with u in tuples). This allows all modifiers present within a phrasal group containing a core finding to be automat-

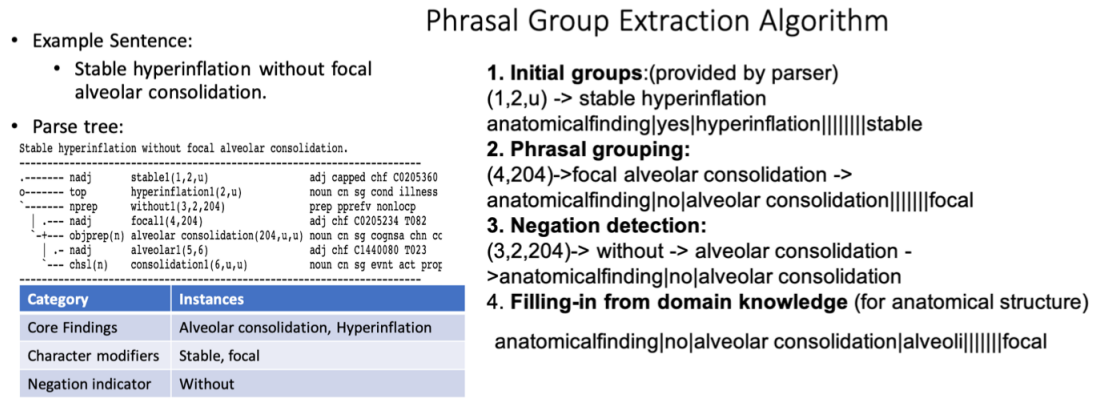

Figure 1: Illustration of phrasal grouping algorithm.

ically associated with the finding. For example, the modifier "stable" is associated with the core finding "alveolar consolidation" in Figure 1. Finally, to determine if a core finding is a positive or negative finding (e.g. "no pneumothorax"), we use a two-step approach that combines language structuring and vocabulary-based negation detection (Guo et al., 2017). The negation pattern detection algorithm identifies words within the scope of negation by iteratively expanding neighborhood of seed negation terms by traversing the dependency parse tree of a sentence (Guo et al., 2017).

3.3. Self-supervised Clinical Similarity Learning

In this paper, we develop a new embedding that is designed to capture similarity in the meaning and sense of multiword clinical terms. It uses a supervised contrastive learning approach to build an encoder for each clinical term w_i in the vocabulary V using the initial unordered similarity lists provided by an ontology such as the chest X-ray lexicon $S_i = \{w_j | w_j \in V$ and w_j is clinically synonymous with $w_i\}$. The similarity lists obtained for each clinical term vary in size from 1 to as large as 4,000 in the 11,000 term chest X-ray lexicon. Note that due to the sub and super concepts being present in the similarity lists, these lists are not necessarily symmetrical and cannot be recursively merged to form larger groups without diluting the underlying semantics. Table 3 shows examples of such similarity lists for a few clinical terms.

We now develop an embedding that captures the essence of these similarity lists and puts an ordering metric to allow objective comparisons during search. It pulls together all members of the similarity list of an anchor clinical term as positive samples and pushes apart the rest as negative examples using a loss designed for this purpose originally for image classification (Khosla et al., 2020). Given a clinical term w_i, we encode it by a one-hot encoding $I_i \in \{0,1\}^{|V|}$, $s.t. \sum_{i=1}^{|V|} I_{ij} = 1$ as an input to the network. As a supervision label, we form a binary label vector $Y_i \in \{0,1\}^{|V|}$, $s.t. \sum_{i=1}^{|V|} Y_{ij} = 1$ iff $w_j \in S_i$ and 0

otherwise and $\sum_{j=1}^{|V|} Y_{ij} = |S_i|$. Thus each similarity list is characterized by a unique binary pattern label vector.

We generate a new encoder-decoder network consisting of an encoder as a dense fully connected layer with ReLU activation and a decoder/projection network as another fully connected layer with ReLU activation. The encoder maps I_i to a representation vector R_i normalized to unit hypersphere, and the projection network renders the output z_i to match the expected binary pattern vector Y_i. The similarity between an anchor clinical term w_i at index i in the ordered vocabulary V , and a candidate term w_{ji} that is originating from the same similarity list S_i at index j in the ordered vocabulary V can be captured by the contrastive loss per similarity list as:

$$L_{contrast}(S_i) = \sum_{w_j \in S_i} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \tag{1}$$

Here z_i is the projected vector for word w_i and z_j is the projected vector similarly for w_{ji}. Finally, z_a is the projected vector for any term w_a either inside or outside the similarity list (i.e. ideally the entire vocabulary). In general, since the similarity lists are small in size, the number of negative samples to differentiate them need not take up the entire vocabulary V, so smaller batch sizes could be used. τ is the temperature to weigh the contribution from similar vectors. Also, since there are multiple such similarity lists, one for each vocabulary term, we can train them in sequential fashion through batching using a cumulative contrastive loss as:

$$L_{contrast} = \sum_{j}^{|V|} L_{contrast}(S_j) \tag{2}$$

Overall, the model was trained with 237 similarity lists from the chest X-ray lexicon covering a 11,000 term clinical vocabulary pertinent to chest X-rays. The designed network architecture had the following parameters: input and output vector sizes= 11000, encoding size =300 , temperature=0.05. We tuned the performance varying batch sizes from 200 to 1,000 words, and epochs from 5 to 50. Convergence per similarity lists was usually achieved within 5 to 10 epochs. We used the Adam optimizer for fast convergence with the learning rate as 0.001. Two NVIDIA P100 GPUs with 16 GB were used for training and training took 5 min per batch. The network overall had around 33 million parameters.

3.4. Forming a report index for search

Given a report R_k, and a sentence $S_j \in R_k$, we use the approach described in Section 3.2 to extract one or more FFL patterns F_i from S_j. Let w_l denote a clinical term within F_j. For example, an FFL pattern *disease|yes|calcified granuloma|stable* has 4 clinical terms *disease, yes, calcified granuloma* and *stable*. The second term in particular, will be important to retain in the index to avoid match to negations by design. Then using the similarity encoding above, each such term can be expanded to clinically similar terms as $G_l = \{w_p \in V | d(W_l, W_p) < \delta\}$ where W_p, W_l are the similarity encoding vectors of the terms w_p, w_l respectively and $d(:)$ is the cosine distance between the encoding vectors. Using each of the

w_p we can now create a reverse index entry $w_p -> (F_i, S_j, R_k)$. Thus each sentence will be indexed by multiple terms w_p corresponding to clinically synonymous word variants of elements in their constituent FFL patterns which already group all elements of a concept found within a sentence. This pre-processing trades off search time for storage size in the index. By pre-indexing for all synonymous term variants, we handle term variants of queries. We also ensure high sensitivity or recall while maintaining search as an O(1) operation per lookup. It also avoids the need for query expansion during search.

3.5. Searching reports in response to queries

Given a query q consisting of a set of clinical terms, a similar process to that described in Section 3.2 can be used to form FFL patterns F_q from the query as well by using the lexicon vocabulary. Here we assume that all queries can be formulated using one or more clinical terms from the vocabulary within the lexicon or those terms will be not be used during search and encoding formation. Using each clinical term $w_q \in F_q$, we can simply lookup the report index to find all possible matching FFL patterns from the report sentences. By accumulating the matches per term, we form a histogram of hits $h(F_i)$ for all matching $F_i \in S_j \in R_k$. Since the FFL patterns contain a minimum of 3 clinical terms (concept type, positive or negative finding, core finding), retaining $h(F_i) > 3$ ensures high recall. However the precision can still be low with such an approach as a large number of matches are still possible. In our approach we ensure $|h(F_i) - |F_q||$ is as low as possible implying the FFL patterns with the highest value of $h(F_i)$ will be retained in this step.

3.6. Final ranking using use context with BERT

The index lookup and ranking step in the previous step pruned a large number of candidates from the report index while still ensuring a high recall. In the final step, we prune the resulting ranked list using transformer methods to encode the overall clinical use context captured in the FFL patterns. Specifically, we form an average vector from the query FFL pattern $B(F_q) = \sum_{i=1}^{n} B(w_{qi})$ where $B(w_{qi})$ is the BERT-encoded vector for word $w_{qi} \in F_q$. The $B(F_i)$ can similarly be constructed from each of the ranked FFL patterns matched. Cosine distance between the BERT query phrase vector and those of the FFL patterns is used to produce the final ranking and a threshold is chosen. Since the order of modifiers is fixed in the FFL pattern, the local word context in the FFL pattern is now well captured by transformer-based methods (Huang et al., 2019; Lee et al., 2020). As the precision and recall vary based on the chosen threshold, a cross-validation analysis was done to choose an optimal threshold to balance between precision and recall. While our implementation currently use the pre-trained BERT from Huggingface (bertbase-uncased), other variants of BERT could be used including clinicalBERT (Huang et al., 2019).

4. Cohort

To test the efficacy of our approach we formed an evaluation cohort of report dataset and queries as described below.

4.1. Dataset

Our experiments were carried out on a public collection of radiology reports provided by Indiana University (Demner-Fushman et al., 2016). The institutional clearances as well as the inclusion and exclusion criteria for forming this publicly available dataset has already been described in (Demner-Fushman et al., 2016). Specifically, a total of 2,557 unique reports were found in this collection after pruning for duplicates and 10,980 sentences were extracted from the reports. Using the FFL pattern extraction algorithm, a total of 17,174 patterns were extracted from these sentences averaging 1.56 FFL patterns per sentence. By pooling all clinical terms within FFL patterns, we obtained 80,122 clinical terms. The whole index creation process for this report collection took less than 10 seconds to generate. With larger report collections, this operation can be made scalable by recording the index in a commercial search engine such as ElasticSearch (Gormley and Tong, 2015).

4.2. Feature choices

Each clinical term found in any FFL pattern was represented by an embedding feature vector using similarity encoding and neighborhood explorations in the embedding space to retrieve all related embedding vectors. For our lexicon, this generated 2,952,578 unique terms that were then used to create the report index.

4.3. Query selection

In order to test the features of different search algorithms, we formed a set of disease queries by selecting the names of diseases found in the Indiana collection that had at least one report occurrence. To each of these we added select modifiers characterizing location, laterality, severity, hedging, and co-association as found commonly in reports. This list is shown in Table 4 along with the number of report occurrences in our collections that constitute a match to the query. As can be seen, some of the queries are semantically equivalent and their results are expected to be identical (for example, "granuloma" and "granulomatous disease"). A total of 42 queries were used to test all search methods.

4.4. Ground truth evaluation

For each of the queries, a set of ground truth matching sentences were recorded from the Indiana reports through manual inspection by 2 domain experts. The resulting ground truth database recorded 945 entries of matches to all 42 queries recorded with a key that captured both report IDs and sentence IDs per query. Since queries are not likely to ask for missing diseases rather than presence of diseases, we only considered positive instances of diseases in our queries even though the report index itself captured negative occurrences of concepts.

5. Results

In this section, we present our results of testing the proposed approach on the benchmark queries and the associated report dataset.

Query	N	Query	N
calcified granulomatous disease	137	fluid overload	10
airspace disease	117	interstitial lung disease	9
prior granulomatous infection	77	penumonia	8
old granuloma	77	atypical infection	3
previous granuloma	77	perihilar calcified granulomas	3
copd or emphysema	51	pericardial effusion	3
emphysema	51	nodules compatible with granuloma	2
copd	51	early pneumonia	2
unchanged granuloma	44	right upper lobe penumobia	1
stable granuloma	44	chest infection	1
right lower lobe granuloma	24	bilateral copd	1
granulomatous disease	20	prior asbestos exposure	1
granuloma	20	mostly likely copd	1
bilateral granuloma	17	pulmonary edema	1
aortic calcifications	14	small granulomatous disease	1
granuloma in the left lower lobe	13	opacities with infection	1
granuloma in the left upper lobe	13	bilateral penumonia	1
osteopenia or other skeletal disease	12	cirrhosis	1
skeletal diseases	12	suspected emphysema	1
consistent with emphysema	11	hyperexpansion consistnet iwth emphysema	1
chronic granuloma	10	aveolar hemorrhage	1

Table 4: Illustration of queries and the corresponding number (N) of ground truth matches to test various search algorithms.

5.1. Evaluation metrics

For all the methods tested, we evaluated for each query, a per query precision as the ratio of number of matching keys over the total number of retrieved keys, while per query recall was recorded as the ratio of matching keys to the size of the ground truth match set for the query. In addition, for each query, we recorded the rank of the match for each key in the ground truth list for the query to compute the mean average precision as described in (map).

5.2. Comparison methods

To compare our approach to state-of-the-art algorithms for medical text retrieval, we implemented a classical Term Frequency and Inverse Document Frequency (TF-IDF) information retrieval method that creates a term-document matrix. TF-IDF is made up of two parts, term frequency (TF) and inverse document frequency (IDF). TF gives the number of times a term occurs in a document. IDF is computed as the inverse frequency of documents containing the searched term. TF-IDF rewards term frequency and penalizes document frequency. BM25 (Robertson and Zaragoza, 2009)improves upon TF-IDF by accounting for document

Method	Reports	Sent.	Queries	Avg.Prec.	Avg.Recall	MAP	Total Time
Ours	2,557	17,174	42	0.41	0.91	0.43	1.62s
BM25	2,557	17,174	42	0.28	0.53	0.33	0.05s
SBERT	2,557	17,174	42	0.09	0.13	0.11	10.50s

Table 5: Retrieval performance comparison of different methods and ablation studies.

length and term frequency saturation. BM25 represents state-of-the-art TF-IDF-like retrieval functions in document retrieval, and often achieves better performance compared to TF-IDF. Hence we adopted the BM25 implementation for a comparison method based on information retrieval approach.

To compare with neural network approaches, we adopted sentence-BERT (SBERT) (Reimers and Gurevych, 2019). In SBERT, a siamese network architecture is used to embed queries and documents into the same latent space, enabling nearest-neighbor semantic retrieval. In inference, since the representation of the candidate texts can be pre-computed, only the query embedding needs to be computed, enabling fast and efficient retrieval. We used an identical network to encode the queries and the sentences. The Clinical BioBert (Huang et al., 2019) was used as our encoder, which was initialized from BioBERT (Lee et al., 2020) and trained on clinical text. The MEAN-pooling strategy was used the generate the fixed sized sentence embeddings (by computing the mean of all output vectors). We employed the same rationale in (Yang et al., 2019; Zhang et al., 2020), which assumes that the "best" sentence in a document provides a good proxy for document relevance. Therefore, in retrieval, we compute the cosine similarity between the query and every sentence in the patient reports, and select the highest sentence score as the report score for ranking.

5.3. Performance

The results of our comparison analysis is shown in Table 5. As can be seen, our method yields far higher average recall than the comparable methods while still achieving more than 27% improvement in precision over the nearest comparison. The increased recall is due to the indexing step that already incorporates synonym variants, and fine-grained description in the FFL pattern. The increased precision is due to the pruning of negation matches and the computation of average vectors in BERT encoding from clinical terms in the FFL pattern rather than the entire sentence. Although our method is currently slower than the BM25, it is amenable to scaling with fast search engine implementations when the report collection grows large due to the use of an index.

6. Discussion

In this paper, we have introduced several enhancements compared to a traditional IR or neural IR approaches in searching for clinically meaningful matches to queries in documents. First, our representation of document text is in the form of clinical context vectors derived automatically as FFL patterns. Secondly, we introduce a new similarity preserving contrastive embedding to capture all meaning-wise variants and spelling variants of concepts. Finally, our search method allows for fast lookup (O(1)), while still allowing for synonym expansions and term variants.

It is a hybrid approach combining principles of information retrieval and textual encoding approaches with the natural language analysis of sentences in reports for fine-grained descriptors of concepts. The new similarity-preserving encoding leads to a marked improvement in average recall performance across queries as well as significant improvement in mean average precision.

In a clinical use case, the implications of using our search engine would be greater flexibility in querying while still ensuring high sensitivity and specificity in capturing the intentions of the user.

Note that the neural vector search of lexicon terms derived from contrastive embedding is still a better alternative to direct lookup of lexicon where a small variation in the query could lead to a miss in the lookup altogether.

Limitations Our approach, however, currently relies on the availability of a domain lexicon preferably composed by clinicians who can reflect an understanding of clinical concepts and their equivalent terms. Automatic generation of such lexicons for different domains is an interesting future research direction. Further, the current multi-step approach could be optimized to develop an end-to-end machine learning module that incorporates all the constituent processing, an aspect that will be explored in future.

References

Introduction to information retrieval. URL https://web.stanford.edu/class/cs276/handouts/EvaluationNew-handout-1-per.pdf.

Roman Bartusiak, Łukasz Augustyniak, Tomasz Kajdanowicz, Przemysław Kazienko, and Maciej Piasecki. Wordnet2vec: Corpora agnostic word vectorization method. *Neurocomputing*, 326:141–150, 2019.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Clinton Gormley and Zachary Tong. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine.* " O'Reilly Media, Inc.", 2015.

Yufan Guo, Deepika Kakrania, Tyler Baldwin, and Tanveer Syeda-Mahmood. Efficient clinical concept extraction in electronic medical records. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Yufan Guo, Joy Wu, Tyler Baldwin, David Beymer, Vandana V Mukherjee, and Tanveer F Syeda-Mahmood. Improving the path from diagnoses to documentation: A cognitive review tool for clinical notes and administrative records. In *AMIA Annual Symposium Proceedings*, volume 2018, page 518. American Medical Informatics Association, 2018.

David A Hanauer, Qiaozhu Mei, James Law, Ritu Khanna, and Kai Zheng. 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). *Journal of biomedical informatics*, 55:290–300, 2015.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Ronilda Lacson, Nathanael Sugarbaker, Luciano M Prevedello, Ivan IP, Wendy Mar, Katherine P Andriole, and Ramin Khorasani. Retrieval of radiology reports citing critical findings with disease-specific customization. *The open medical informatics journal*, 6:28, 2012.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Michael C McCord, J William Murdock, and Branimir K Boguraev. Deep parsing in watson. *IBM Journal of research and development*, 56(3.4):3–1, 2012.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144, 2008.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Karthik Natarajan, Daniel Stein, Samat Jain, and Noémie Elhadad. An analysis of clinical queries in an electronic health record search utility. *International journal of medical informatics*, 79(7):515–522, 2010.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

Irena Spasic, Dominik Krzeminski, Padraig Corcoran, and Alexander Balinsky. Cohort selection from longitudinal patient records: Text mining approach. *JMIR Medical Informatics*, 7, 10 2019. ISSN 22919694. doi: 10.2196/15980.

Irena Spasic, Goran Nenadic, et al. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984, 2020.

Tanveer Syeda-Mahmood. Similarity retrieval of cardiac reports. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 1135–1141. IEEE, 2010.

Tanveer Syeda-Mahmood, Ken CL Wong, Joy T Wu, Ashutosh Jadhav, and Orest Boyko. Extracting and learning fine-grained labels from chest radiographs. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1190. American Medical Informatics Association, 2020.

Joy T Wu, Ali Syed, Hassan Ahmad, Anup Pillai, Yaniv Gur, Ashutosh Jadhav, Daniel Gruhl, Linda Kato, Mehdi Moradi, and Tanveer Syeda-Mahmood. Ai accelerated human-in-the-loop structuring of radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1305. American Medical Informatics Association, 2020.

Wei Yang, Haotian Zhang, and Jimmy Lin. Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*, 2019.

Haotian Zhang, Gordon V Cormack, Maura R Grossman, and Mark D Smucker. Evaluating sentence-level relevance feedback for high-recall information retrieval. *Information Retrieval Journal*, 23(1):1–26, 2020.