Improving Neural Models for Radiology Report Retrieval with Lexicon-based Automated Annotation

Anonymous ACL submission

Abstract

Many clinical informatics tasks that are based on electronic health records need relevant patient cohorts to be selected based on findings, symptoms, and diseases. Frequently, these conditions are described in radiology reports which can be retrieved using information retrieval (IR) methods. The latest of these techniques utilize neural IR models such as BERT trained on clinical text. However, these methods still lack semantic understanding of the underlying clinical conditions as well as ruled out findings, resulting in poor precision during retrieval. In this paper we combine clinical finding detection with supervised query match learning. Specifically, we use lexicon-driven concept detection to detect relevant findings in sentences. These findings are used as queries to train a Sentence-BERT (SBERT) model using triplet loss on matched and unmatched query-sentence pairs. We show that the proposed supervised training task remarkably improves the retrieval performance of SBERT. The trained model generalizes well to unseen queries and reports from different collections.

1 Introduction

Electronic health record (EHR) retrieval is important for clinicians, staff and researchers. The tools for performing clinically relevant searches could aid in many use cases such as clinical decision support (Syeda-Mahmood, 2010), auditing, revenue cycle management, and cohort selection for clinical studies. Frequently, these searches involve retrieval of patients based on clinical findings that are often captured in unstructured textual reports such as radiology reports, encounter notes, etc. Unlike structured query-based lookup of EHR, retrieval of unstructured (free-text) EHRs is much more challenging, requiring a semantic understanding of the underlying clinical conditions present or absent. Conventional exact or approximate term-based retrieval methods such as BM25 (Robertson and Zaragoza, 2009) often perform poorly in response to ad-hoc queries (Chamberlin et al., 2020), as these methods lack the ability of semantic understanding of the clinical as well as language context. With the emergence of deep learning encoding models, new retrieval methods have emerged with studies showing BERT-based neural methods outperforming BM25 on multiple retrieval benchmarks (Yilmaz et al., 2019a; Chang et al., 2020; Nogueira and Cho, 2019; Yilmaz et al., 2019b; Qiao et al., 2019). The BERT-based retrieval methods can be classified into two categories: the cross-attention (or interaction-based) models (Yilmaz et al., 2019a; Nogueira and Cho, 2019; Yilmaz et al., 2019b) and the embedding-based (or representation-based) models (Chang et al., 2020; Reimers and Gurevych, 2019). While the BERT-style cross-attention models are very successful, they cannot be directly applied to large-scale retrieval problems because computing the similarity score for every possible query-document pair during inference can be prohibitively expensive. Therefore, they were often used as a re-ranker after a initial candidate retrieval round using BM25. The embedding-based methods can pre-encode the documents, and only the queries need to be encoded upon retrieval. Retrieval can be achieved via approximate nearest-neighbor search in the embedding space very efficiently (Johnson et al., 2021). In this study, we focus on the embedding-based retrieval BERT models. Specifically, we adopted the sentence-level retrieval setting, as studies suggested that the "best" sentence in a document provides a good proxy for document relevance (Yilmaz et al., 2019a).

Different pre-training tasks were used to train the BERT-based models for retrieval. The pre-training tasks range from masked language modelling (MLM) over unlabeled free-text to supervised training on labeled datasets such as STS (Cer et al., 2017), MS MARCO (Nguyen et al., 2016) or TREC Microblog track (Lin et al., 2014). How-

ever, MLM is not tailored for the purpose of information retrieval (IR), and labeled datasets are usually small and not easily accessible. Recently, pre-trained models on biomedical corpora such as BioClinicalBERT (Alsentzer et al., 2019) and BioBERT (Lee et al., 2020) can obtain embeddings with medical-domain-specific knowledge, but they were still trained with MLM.

Early studies (Natarajan et al., 2010) showed that most clinical queries are actually short queries (e.g. a disease or a syndrome). We found that the existing BERT models pre-trained with MLM performed poorly on short queries as well as negative queries (i.e. queries asking for lack of a finding). Ideally, if retrieval systems could be trained by matched and unmatched query-sentence pairs, in both positive and negated instances, we can expect a higher precision and recall in retrieval. However, manually labeling a large dataset is impractical, particularly for the medical domain where the number of clinical findings is very large. Training neural IR models using weak supervision has been previously investigated (Dehghani et al., 2017; MacAvaney et al., 2019), which use unsupervised methods (e.g. BM25) or article headings to provide pseudo labels. However, these pseudo labels are usually imprecise and the article headings are not always available.

Motivated by these challenges, we present a hybrid approach where we combine automated clinical finding detection with supervised query-sentence pair learning. Specifically, we use an automatic lexicon-driven concept detection method to detect relevant chest X-ray (CXR) findings in sentences. These findings paired with the sentences containing them serve as weakly labeled training data for Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). The resulting approach avoids manual annotation and can be scaled for training on a large number of query-sentence pairs. We show that the proposed training task remarkably improves the retrieval performance of SBERT on datasets with automatic annotations and human annotations.

2 Methods

2.1 Fine-grained concept extraction

The algorithm for extracting findings from sentences in reports uses a vocabulary-driven approach. Specifically, a domain-specific CXR finding lexicon was used. This lexicon captures the name of finding along with its potential variants and synonyms mined from over 200,000 chest radiology reports. To spot the occurrence of a finding lexicon phrase within reports, a string matching algorithm called the *longest common subfix (LCF)* algorithm was used. To determine if a core finding is positive or negative (e.g. "no pneumothorax"), a two-step approach that combines language structuring and vocabulary-based negation detection is used. The method is reported to be highly accurate (<3% errors) compared with human labels. More details are described in (Syeda-Mahmood et al., 2020).

2.2 Labeled data generation

In this paper, we focus on "anatomical findings" as well as "disease concepts" as those are the most commonly searched in EHR (Natarajan et al., 2010). We use these finding modifiers as surrogates for queries. For each sentence S_j in our data collection, we have a set with K_j labeled data entries $I_j = \{(S_j, N_{j,i}, M_{j,i})\}_{1 \leq i \leq K_j}$. For each labeled entry $(S_j, N_{j,i}, M_{j,i})$, $M_{j,i}$ is the i-th finding for S_j, and $N_{j,i} = yes|no$ indicates a positive or ruled out finding. By using the findings as query surrogates, we can designate a query $Q_{j,i} = (N_{j,i}, M_{j,i})$ paired with S_j: if $N_{j,i}$ equals to yes, $Q_{j,i}$ is a positive query, otherwise $Q_{j,i}$ is a negative query. For example, $(yes,\ vascular\ congestion)$ and $(no,\ pulmonary\ edema)$ are two queries for the sentence "lungs: central vascular congestion without overt edema." The actual queries may be more properly phrased such as "presence/absence of X".

Since we labeled all the sentences in our training dataset extensively with all the finding types we summarized, we can create a dictionary using each unique query $Q = (N, M)$ as the key and the list of all the sentences that contain that query as the dictionary value. Any sentence in the list is considered as a matched sentence for that query, whereas other sentences are considered as unmatched sentences.

2.3 Model

We used SBERT as our retrieval model. MEAN-pooling was used to derive a fixed size sentence embedding (for either queries or EHR sentences). We used the triplet objective function (Reimers and Gurevych, 2019) to train our model. Given a query q, a matched sentence m and an unmatched sentence u, the triplet loss tunes the network such that the distance between q and m is smaller than the distance between q and u by a margin ϵ:

$$max(\|e_q - e_m\| - \|e_q - e_u\| + \epsilon, 0) \quad (1)$$

where e_q, e_m and e_u are the sentence embeddings for q, m and u, respectively. $\|\cdot\|$ is a distance metric. We used the cosine distance and $\epsilon = 0.5$.

To improve training, we further used hard-sampling (HS) to mine the hardest unmatched sentence for the triplet loss within a training batch. To be specific, we performed inference within a batch beforehand to find the unmatched sentence with the highest cosine similarity score (the most confusing unmatched sentence) for each query. We further applied mega-batching (MB) (Wieting and Gimpel, 2018) to encourage the model to learn to distinguish "harder" unmatched sentences by increasing the batch size.

3 Experiments and Results

3.1 Datasets

The experiments in Section 3.1-3.4 were carried out on two public collections of radiology reports provided by Indiana University (Demner-Fushman et al., 2016) and NIH (Wang et al., 2017). After pruning for duplicates and applying our labeled data generation algorithm described in Section 2.2, a total of 21,612 labeled entries were generated for the Indiana dataset, which include 10,363 unique sentences, 200 positive queries and 75 negative queries. For the NIH dataset, 17,047 labeled entries were generated, including 9,091 unique sentences, 250 positive queries and 30 negative queries.

3.2 Ablation study and parameter tuning

We first run an ablation study on the Indiana dataset (IND) to investigate if hard-sampling (HS) and mega-batching (MB) can bring improvement over random-sampling (RS, randomly select unmatched sentence within a batch) and normal-batching (NB, size 32). We randomly split the IND dataset into two halves with non-overlapping findings with the constraint that they should roughly have equal number of labeled entries. After the split, the two sets have 117/44 and 83/31 positive/negative queries, respectively. We performed 2-fold cross-validation and reported the average of the two test results regarding mean Average Precision (mAP). This allows us to evaluate the model performance on unseen queries. The evaluation was performed over positive queries (Pos. Q.), negative queries (Neg. Q.) and all queries (All Q.) separately.

The results in Table 1 shows that the combination of HS and MB achieved the best results. Increasing the mega-batching size to 128 resulted the best performance, but further increasing the batch size slightly degraded the performance. The remarkable improvement of SBERT over the baseline BioClinicalBERT also suggests that the proposed model can generalize well to unseen queries.

Model	mean Average Precision (mAP)		
	Pos. Q.	Neg. Q.	All Q.
BioClinicalBERT	0.213	0.254	0.224
SBERT/RS/NB(32)	0.353	0.312	0.349
SBERT/HS/NB(32)	0.384	0.334	0.371
SBERT/HS/MB(64)	0.388	0.318	0.369
SBERT/HS/MB(128)	**0.399**	**0.392**	**0.397**
SBERT/HS/MB(256)	0.392	0.352	0.381
SBERT/HS/MB(512)	0.380	0.344	0.370

Table 1: Ablation study and hyperparameter tuning on the Indiana dataset.

3.3 Cross-dataset study

We also trained on the IND dataset and tested on the unique sentences in the NIH dataset and vice versa to investigate whether a trained model can generalize well to a different dataset. The best SBERT model from Table 1 was used here. We further included Okapi BM25 (k_1=1.5, b=0.75), the pre-trained BERT (Huggingface "BERT-base-uncased"), the fine-tuned BERT (trained on the EHR sentences using MLM, without using our generated annotations), the BioClinicalBERT and SBERT pre-trained on MS MARCO dataset for comparison. More details about these models are given in the appendix. In addition to mAP, mean Recall (over all the queries) was also reported, where Recall was defined as the ratio of the number of correctly retrieved sentences to the size of the query's ground truth list.

Table 2 shows that our fine-tuned SBERT performs very well on the dataset from another collection regarding both mAP and mR, and outperformed the other BERT/SBERT models by large margins. The baseline BERT without pre-training over medical texts obtained the worst results. The results for BERT (fine-tuned) and BioClinicalBERT suggest that MLM training over the texts from the same domain can lead to some improvements but is still not ideal for direct use of retrieval. SBERT pre-trained on MS MARCO dataset showed significant improvements over BERT trained with MLM, but lacks domain-specific knowledge and shows performance drop on negative queries. BM25 performs well on positive queries with performance degradation on negative queries as well, because negation is not always

Model	mean Average Precision (mAP)			mean Recall (mR)		
	Pos. Q. IND / NIH	Neg. Q. IND / NIH	All Q. IND / NIH	Pos. Q. IND / NIH	Neg. Q. IND / NIH	All Q. IND / NIH
BM25	0.39 / **0.46**	0.34 / 0.32	0.38 / 0.44	0.36 / **0.43**	0.30 / 0.27	0.35 / 0.42
BERT	0.14 / 0.16	0.21 / 0.23	0.16 / 0.17	0.12 / 0.15	0.19 / 0.23	0.14 / 0.16
BERT (fine-tuned)	0.20 / 0.23	0.22 / 0.23	0.21 / 0.23	0.19 / 0.21	0.21 / 0.21	0.19 / 0.21
BioClinicalBERT	0.16 / 0.28	0.21 / 0.25	0.17 / 0.27	0.14 / 0.27	0.19 / 0.22	0.15 / 0.26
SBERT (MS MARCO)	0.40 / 0.44	0.35 / 0.36	0.39 / 0.43	0.37 / 0.40	0.31 / 0.31	0.35 / 0.39
SBERT (ours)	**0.48** / 0.45	**0.42 / 0.56**	**0.46 / 0.47**	**0.44** / 0.42	**0.39 / 0.47**	**0.42 / 0.43**

Table 2: Cross-dataset evaluation. The dataset name in the heading means the model was tested on that dataset.

explicitly expressed in EHR.

3.4 Embedding separation analysis

Model	IND	NIH
BERT	-0.04±0.06	0.01±0.07
BERT (fine-tuned)	0.03±0.09	0.05±0.08
BioClinicalBERT	0.01±0.05	0.01±0.03
SBERT (MS MARCO)	0.01±0.01	0.02±0.01
SBERT (ours)	0.42±0.36	0.56±0.34

Table 3: Embedding space separation analysis.

Because we have the negation labels, we can also create opposite-negation queries. For example, the opposite-negation query for "no opacity" would be "opacity". Ideally, for a given sentence, the similarity score between the matched query and sentence should be larger than that between the opposite-negation query and the sentence. We reported (Table 3) the differences (mean±std) between these two scores for all the entries in each dataset with all the BERT embedding-based methods. Our trained SBERT showed a clear separation in the embedding space. The distances for the other BERT models are all around zero with even negative distances, suggesting poor negation awareness.

3.5 Evaluation on human-annotated data

We also evaluated our model on a separate human-annotated dataset. 206 CT reports and 120 CXR reports were annotated by 3 radiologists on sentence-level with majority voting (more details in appendix). This resulted in 2,990 unique sentences/8 queries for CT reports and 1,810 unique sentences/18 queries for CXR reports. Table 4 shows that our SBERT fine-tuned on either IND or NIH dataset outperforms the other compared methods by large margins.

Model	mAP	mR
	CT / CXR	CT / CXR
BM25	0.34 / 0.34	0.39 / 0.31
BioClinicalBERT	0.32 / 0.28	0.34 / 0.26
SBERT (MS MARCO)	0.35 / 0.39	0.40 / 0.37
SBERT (trained on IND)	**0.59 / 0.66**	0.57 / 0.60
SBERT (trained on NIH)	0.52 / 0.57	0.50 / 0.54

Table 4: Evaluation on human-annotated datasets.

4 Discussion

In this paper we demonstrated that the proposed supervised pre-training tasks with automated annotation can greatly improve the IR performance of SBERT on short and negative queries. The proposed labeled data generation method can also be used to train the cross-attention BERT models for further improvement when computation speed is not the bottleneck.

We focused on short queries in this study, and BM25 still performs well on positive queries. The embedding-based BERT models are expected to show more advantages over BM25 on complicated queries that require semantic understanding. Having the comprehensive negation and finding labels for each sentence also allows us to assemble more complicated queries that include more than one finding, such as “A and B” or “A without C” where A, B and C represent three different findings. These more challenging tasks can be explored in the future work. The label generation tool can also be extended to training IR models in domains other than medical domain, such as finance, law, or retail, provided with the corresponding lexicons.

5 Conclusion

In this work we proposed to generate query-sentence pairs automatically using a CXR lexicon for training embedding-based BERT models on the EHR retrieval problem. We showed that the fine-tuned SBERT obtained a substantial performance gain over the other pre-trained models. The trained model can also generalize well to unseen queries and data from another source. The proposed method can be especially helpful in training and evaluating neural IR models in domains with limited human-labeled data.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Steven R Chamberlin, Steven D Bedrick, Aaron M Cohen, Yanshan Wang, Andrew Wen, Sijia Liu, Hongfang Liu, and William R Hersh. 2020. Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. *JAMIA open*, 3(3):395–404.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 65–74, New York, NY, USA. Association for Computing Machinery.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

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

Jimmy Lin, Yulu Wang, Miles Efron, and Garrick Sherman. 2014. Overview of the TREC-2014 microblog track. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, volume 500-308 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019. Content-based weak supervision for ad-hoc re-ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 993–996, New York, NY, USA. Association for Computing Machinery.

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

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the behaviors of BERT in ranking. *CoRR*, abs/1904.07531.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

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

Tanveer F. Syeda-Mahmood, Ken C. L. Wong, Joy T. Wu, Ashutosh Jadhav, and Orest B. Boyko. 2020. Extracting and learning fine-grained labels from chest radiographs. In *AMIA 2020, American Medical Informatics Association Annual Symposium, Virtual Event, USA, November 14-18, 2020*. AMIA.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019a. Applying bert to document retrieval with birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 19–24.

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019b. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3481–3487.

A Appendix: Model training details

Here we provide more details on the models used in Section 3. We used the Huggingface "BERT-base-uncased" model (pre-trained on BookCorpus and English Wikipedia, availabel at: https://huggingface.co/bert-base-uncased) as our BERT model for comparison. The BERT (fine-tuned) model was fine-tuned on the EHR text (Indiana or NIH dataset) using MLM for 5 epochs based on the "BERT-base-uncased" model. The pre-trained BioClinicalBERT (Alsentzer et al., 2019) (availabel at: https://github.com/EmilyAlsentzer/clinicalBERT) was initialized with BioBERT (Lee et al., 2020) and fine-tuned on clinical notes.

Our SBERT model was initialized with the BioClinicalBERT. We fine-tuned SBERT using triplet loss for 10 epochs for all datasets in this study. We used AdamW optimizer with learning rate 2e-5, weight decay 0.01 and a linear learning rate warm-up of 100 steps.

The SBERT model used as comparison was pre-trained on 500K (query, answer) pairs from the MS MARCO dataset. This pre-trained model (msmarco-bert-base-dot-v5) was one of the recommended sentence embedding models from the official SBERT webpage (https://www.sbert.net/docs/pretrained_models.html). Among all the pre-trained models, we picked this one because it is the only pre-trained model based on "BERT-base" model, to be consistent with all the other models (all based on "BERT-base") in our experiments. Since this model was tuned to be used with dot-product, we used dot-product to calculate similarity scores only for this model in the retrieval experiments in Table 2. For all the other models, cosine-similarity was used to calculate scores. However, for the embedding separation analysis in Table 3, cosine-similarity was used for SBERT (MS MARCO) as well so that the scale of the similarity scores is comparable to the others.

B Appendix: Human-annotated data

The radiology reports used in Section 3.5 are private data obtained from our collaborative partners. All private data used were anonymized. HIPPA was fully enforced and all data were handled according to the Declaration of Helsinki. 206 CT reports and 120 chest X-ray (CXR) reports were annotated by 3 radiologists on sentence-level using the brat rapid annotation tool (available at https://brat.nlplab.org/). Majority voting was used to handle disagreements. This resulted in 2,990 unique sentences/8 queries for CT reports and 1,810 unique sentences/18 queries for CXR reports. Note that the candidate sentences for retrieval also include those sentences without any our interested disease findings. For the CT reports, the annotation was based on the presence or absence of 4 diseases (resulting in 8 queries): thoracic aneurysm, abdominal aneurysm, lung nodule and pulmonary embolism. The average number of matched sentences for each query is 42$\pm$33. For the CXR reports, the annotation was based on 10 diseases (resulting in 18 queries, as 2 negative queries do not have the corresponding matched sentences): pulmonary embolism, airspace opacity, lung nodule, emphysema, pneumothorax, abdominal aortic aneurysm, thoracic aortic aneurysm, rib fracture, scapula fracture and spine fracture. The average number of matched sentences for each query is 18$\pm$16. It it worth noting that some of the diseases are not even used as queries in the IND/NIH training data, including thoracic aneurysm, abdominal aneurysm and spine fracture.