HHS Public Access

Author manuscript

Int J Med Inform. Author manuscript; available in PMC 2023 May 06.

Published in final edited form as:

Int J Med Inform. ; 158: 104628. doi:10.1016/j.ijmedinf.2021.104628.

Fine-Grained Spatial Information Extraction in Radiology as Two-turn Question Answering

Surabhi Datta, **Kirk Roberts***

School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX

Abstract

Objectives—Radiology reports contain important clinical information that can be used to automatically construct fine-grained labels for applications requiring deep phenotyping. We propose a two-turn question answering (QA) method based on a transformer language model, BERT, for extracting detailed spatial information from radiology reports. We aim to demonstrate the advantage that a multi-turn QA framework provides over sequence-based methods for extracting fine-grained information.

Methods—Our proposed method identifies spatial and descriptor information by answering queries given a radiology report text. We frame the extraction problem such that all the main radiology entities (e.g., finding, device, anatomy) and the spatial trigger terms (denoting the presence of a spatial relation between finding/device and anatomical location) are identified in the first turn. In the subsequent turn, various other contextual information that acts as important spatial roles with respect to a spatial trigger term are extracted along with identifying the spatial and other descriptor terms qualifying a radiological entity. The queries are constructed using separate templates for the two turns and we employ two query variations in the second turn.

Results—When compared to the best-reported work on this task using a traditional sequence tagging method, the two-turn QA model exceeds its performance on every component. This includes promising improvements of 12, 13, and 12 points in the average F1 scores for identifying the spatial triggers, Figure, and Ground frame elements, respectively.

Discussion—Our experiments suggest that incorporating domain knowledge in the query (a general description about a frame element) helps in obtaining better results for some of the spatial and descriptive frame elements, especially in the case of the clinical pre-trained BERT model. We further highlight that the two-turn QA approach fits well for extracting information for complex schema where the objective is to identify all the frame elements linked to each spatial trigger and

*Corresponding author kirk.roberts@uth.tmc.edu (Kirk Roberts).

Author Statement

SD and KR conceived the methodology. SD implemented the system, carried out the experiments, and drafted the initial manuscript. KR supervised the study and acquired funding. SD and KR edited the manuscript and approved the final version.

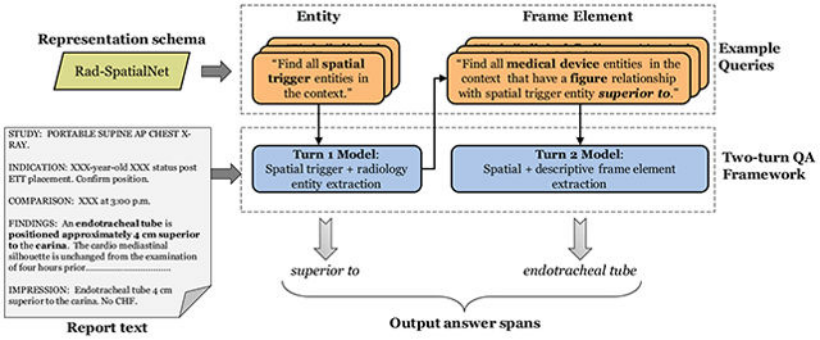
Conflicts of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

finding/device/anatomy entity, thereby enabling the extraction of more comprehensive information in the radiology domain.

Conclusion—Extracting fine-grained spatial information from text in the form of answering natural language queries holds potential in achieving better results when compared to more standard sequence labeling-based approaches.

Graphical Abstract



Keywords

Information Extraction; Spatial Information; Radiology Report; Natural Language Processing; Question Answering; Deep Learning

1. Introduction

Extracting important and detailed descriptions of radiographic findings from reports has lately been of great interest to researchers focusing on natural language processing (NLP) applications in radiology [1-4]. Some primary use cases of the NLP-generated labels include training deep image classification models to predict fine-grained diagnosis [1, 5, 6], phenotyping [7], and automated tracking of findings [8, 9]. Traditionally, sequence classification methods such as conditional random fields [10, 11] and bidirectional Long Short-Term Memory networks [12] have been adopted for extracting information from clinical text including radiology reports. However, because of the availability of well-developed machine reading comprehension (MRC) models, recent research has focused on formalizing information extraction (IE) as question answering (QA) [13-15], highlighting the many advantages of a QA framework over traditional approaches for IE tasks. Additionally, several studies [16-18] explored a new paradigm by formulating relation and event extraction (EE) tasks as multi-turn QA, an approach that involves performing multiple turns of MRC successively. Motivated by these studies, we propose a two-turn QA approach for identifying fine-grained spatial and descriptor information of radiographic findings and medical devices described in radiology reports. To our knowledge, this is the first work in radiology as well as spatial IE that employs a multi-turn QA approach for granular IE.

Radiographic findings and medical devices are usually described by radiologists with reference to some anatomical structures, thus there exists spatial relations between findings/devices and anatomies often connected through spatial phrases. There are also mentions of

other clinically relevant contextual details associated to the spatial relations such as potential diagnoses and a device's distance from the anatomical structure. Moreover, there are spatial and other descriptors describing a radiological entity (e.g., finding, anatomy) that enhance the richness of the labels for the corresponding medical images. The spatial descriptors represent both spatial (e.g., laterality, size, morphology) and other properties of an imaging observation (e.g., composition, distribution pattern, density) described in reports. Other descriptors include status, quantity, temporality, and negation. All these clinically important information are organized according to a frame semantics representation–Rad-SpatialNet proposed in Datta et al. [4]. In frame semantics, a lexical unit (LU) is the word or phrase that invokes a frame and the participants of a frame constitute the frame elements (FEs). In the context of spatial relation frames, an LU is either a spatial preposition/verb (which we refer to as a "trigger") or a radiological entity, and all the spatial roles and descriptors linked to the LU form the FEs. In this work, we refer to the spatial roles (connected to a spatial trigger) and the spatial descriptors (connected to a radiological entity) as "Spatial Frame Elements" (SFEs). The other entity-specific descriptors are referred to as "Descriptive Frame Elements" (DFEs). Creating fine-grained labels requires identifying a wide variety of relations (represented through FEs), some are illustrated in Figure 1, and we focus on extracting such detailed relations in this work.

Recently, both in the general and biomedical domains, studies [14, 19, 20] have demonstrated the effectiveness of formulating named entity recognition (NER) as MRC instead of the traditional sequence labeling technique. Apart from NER, prior work has also framed relation [13] and EE [15] tasks as MRC. Moreover, MRC models have been utilized in a multi-turn QA setting for joint entity-relation extraction [16] and for both general-domain [17] and biomedical [18] EE. Advantages of framing extraction tasks as MRC include leveraging prior knowledge through queries, jointly modeling entities and relations in the form of natural language questions, and making use of advanced MRC models. In addition, multi-turn QA captures the hierarchical dependency of entities and is therefore suitable for complicated scenarios where extraction of certain entities depends on previously extracted entities. As shown in Figure 1, identification of SFEs (e.g., *hemorrhagic foci*) depends on extracting the associated spatial trigger *in*. Similarly, extracting elements such as *enhancing* and *right* are entity-specific. Previous approaches have formulated spatial trigger and element (or role) extraction as a sequence labeling task either in a pipelined or joint learning fashion [4, 21]. However, inspired by the advantages that multi-turn QA provides, we propose to adopt a two-turn QA technique by harnessing MRC models for fine-grained spatial IE from radiology text. A high-level overview of our approach is shown in Figure 2. As can be seen in this figure, framing our IE task as two-turn QA as opposed to single-turn QA is better realized as the FEs are described at the level of each spatial trigger or radiological entity.

In this paper, we aim to extract a comprehensive set of relations pertaining to the common spatially-grounded radiological entities that are of significance to facilitating automated image diagnosis. For this, we frame the problem as two-turn QA where spatial triggers and the main radiological entities (e.g., clinical finding, anatomy) are identified in the first turn and all the SFEs and DFEs (e.g., status descriptors of a finding) in the second. We extract answer spans from the report text by answering template questions (in the form of

natural language queries) constructed separately for entity and FE extraction. The query for the second turn includes the trigger or the radiological entity extracted in the first turn. This makes the query for FE extraction more informative. We experiment with two query variations for extracting the FEs. Our QA framework is based on the pre-trained language model BERT [22]. We compare the performance of our BERT-based two-turn QA approach for extracting spatial information with a sequence labeling approach [4] used as a baseline. Our main contributions are enumerated as follows:

1. Frame the task of spatial IE as two-turn QA.
2. Demonstrate the advantages of applying a QA-based approach over a traditional sequence labeling method for extracting spatial information.
3. Extract more comprehensive information from multiple types of radiology reports targeted toward fine-grained medical image labeling.

2. Materials and Methods

2.1. Data

We use 400 radiology reports from MIMIC-III [23] to extract fine-grained information of radiographic findings and medical devices from three types of reports: chest X-rays, brain MRIs, and babygrams. The entities and frame elements are described in Tables 1 and 2, respectively. Example frame element relations from the dataset are illustrated in Figures 1 and 2. More details can be found in prior work [4].

2.2. Problem Formulation

We formulate the spatial IE problem as a machine comprehension problem where information is extracted from a given text (treated as a context paragraph) using templates posed as queries to elicit specific information (triggers and FEs). The answer spans returned by the MRC system are treated as the extracted entities. In case the system returns a special token NONE, this indicates that the specific entity that is queried for is not present in the report text. Analogous to how a conventional entity-relation extraction system is employed, i.e., first identifying the target entity and then identifying the related entities, our MRC formulation is also designed in two turns/steps as one-time QA is not sufficient to capture this dependency in the information extracted. The target entities extracted in the first turn are mentioned in Table 1. These entities cover a wide range of common radiology terms curated as part of the radiology lexicon, RadLex [24] (see Table 1). The second turn identifies the SFEs associated with a spatial trigger (e.g., Figure, Ground, Diagnosis) as well as spatial (e.g., Laterality) and descriptive (e.g., Status) FEs associated with a radiological entity.

2.3. Query Construction

Entity and frame element (FE) type modification.—The entity and FE types are used in forming the queries. The entity types except ‘Spatial Trigger’, ‘Location Descriptor’, and ‘Quantity’ are modified, as shown in Table 3, while incorporating in a query. We modified the entity types to incorporate more information about the entities in the queries as well as to make the queries sound more natural. The FE types (corresponding to element names)

are used in the queries of the second turn without modification except for the 'Diagnosis' element in which case it is modified to 'Potential diagnosis'.

Target entity extraction.—The modified entity type (of a spatial trigger and a main radiological entity) is converted to a query using a template. This query variant is referred to as $\text{Query}_{\text{find}}$ (shown in Table 4).

Spatial and descriptive frame element extraction.—Each FE is converted to a query, $\text{Query}_{\text{find}}$ (see Table 4), such that the query asks for identifying the text span (belonging to a specific entity type) from the report that has the particular FE relation to a target entity type. The query template contains a slot corresponding to the target entity type (ENT_1) that is filled by the previously extracted entity (ENT_1_SPAN) from the first turn. Thus, this query jointly extracts the FE relation (REL) as well as the related entity (of type ENT_2) in the form of an answer span that is predicted by the MRC model. Using this template, queries are formed such that all FE relations are covered for all possible pairs of target and related entity types. For example, "*find all* ***medical device*** *entities in the context that have a* ***figure*** *relationship with* ***spatial trigger*** *entity* ***above***." is the query constructed for the triplet {spatial trigger, Figure, medical device} where spatial trigger is ENT_1, Figure is REL, and medical device is ENT_2 that is extracted by answering this query. If the answer is NONE, this means there is no such related entity in the text that is associated to the target entity through REL.

We also experiment with another query variation for the second turn. In this, we encode domain knowledge in the query by incorporating a general description of the FE. That is, we prepend a description of the SFE or DFE at the beginning of a query. We refer to this query variation as $\text{Query}_{\text{find + desc}}$ (see Table 4 for template and example). The descriptions developed for each of the FEs are listed in Supplementary Table 1.

2.4. MRC Framework

The MRC architecture is based on the pre-trained language model BERT [22]. Previous work achieved promising results using BERT-based MRC models for QA [22, 25, 26]. We select this model framework owing to the promising performance of using BERT for QA as well as to tackle multi-answer QA. We follow the standard format to feed input into the BERT model for answering queries. We split the whole content of a radiology report into overlapping passages by sliding window and use each passage as context c into the BERT model after combining with the query q. This sliding window technique proved to be effective as evidenced by prior work [16, 27]. After WordPiece tokenization of both query and context, we merge the query q and the context c as [[CLS] q [SEP] c [SEP]] to construct the input sequence where [CLS] and [SEP] are special BERT tokens. As explored in previous work [14, 16], the span extraction mechanism enables queries that have multiple answers given the context passage. Traditional approaches strategize span extraction as two n-class classification problems where one classifier predicts the start index and the other predicts the end index from all the context tokens (n refers to the length of the context passage). However, this strategy is only applicable for single-answer QA settings. To overcome this shortcoming, the two n-class classification task is converted to n 5-class classifications where the softmax function is applied to each token in the context to predict

a BMESO (B-begin, M-middle, E-end, S-single, O-outside) label. This is suitable to our problem where there can be multiple entities of the same spatial or descriptor role that are associated to a single spatial trigger or a radiological finding (see Figure 3). The BERT models for both target entity (turn 1) and FE extraction (turn 2) are trained jointly. The MRC framework and the training mechanism are adopted from a previous work [16].

2.5. Sequence Labeling Baseline

We compare our approach to the BERT-based sequence labeling approach proposed in Datta et al. [4]. In that paper, a $\text{BERT}_{\text{LARGE}}$ model pre-trained on MIMIC-III is fine-tuned to first extract all the spatial triggers in a sentence and then extract the spatial FEs associated with each trigger. A report sentence is represented as [[CLS] *sentence* [SEP]] to feed into BERT to identify the triggers and FEs. Additionally, while extracting the spatial FEs, we mask the spatial trigger identified in the first step to better encode the position of the specific spatial trigger in a sentence for which the FEs are to be identified. The encoder output is then fed into a linear classification layer to predict labels per token. The BIO (B-begin, I-inside, O-outside) scheme is used to tag the triggers and the FEs.

2.6. Experimental Settings and Evaluation

We experiment with both cased and uncased $\text{BERT}_{\text{LARGE}}$ variants in the MRC framework (referred to as Uncased and Cased hereafter). Additionally, we also experiment using a $\text{BERT}_{\text{LARGE}}$ cased version that is pre-trained on MIMIC-III clinical notes for 300K steps [28] (referred as MIMIC+Cased). The hyperparameters used in our experiments are selected based on the validation set and are shown in Table 5. For training the MRC model in the second turn, we only consider the relationships between target and related entities where there is at least one instance of such a relationship in the training data.

We perform 10-fold cross-validation (CV) for evaluating our MRC approach for spatial IE. For each of the 10 iterations, we split the dataset such that reports in 8 folds are used for training and 1 fold each are used for development and testing. We report the average F1 measures for extracting the FEs. Since the query format is the same for the first turn (i.e., target entity identification) and only varied in the second turn (corresponding to using $\text{Query}_{\text{find}}$ and $\text{Query}_{\text{find + desc}}$), we report the average of the two 10-fold CV runs for target entity extraction. We use exact match to evaluate the performance of the MRC approach for both target entity and FE extraction on the test splits. Exact matches of both the target and the related entity spans are required to consider a FE relation extraction as a true positive. We compare our approach to the baseline method for identifying spatial triggers and SFEs connected to triggers. For a fair comparison, we use the same fold settings for 10-fold CV for both the MRC and the baseline methods. The baseline method is also evaluated using exact match for spatial trigger and SFE extraction.

3. Results

The average F1 measures of 10-fold CV evaluation for extracting the spatial and descriptive FEs are shown in Table 6. This includes the results for both the query variations. The average F1 scores of the BERT-based sequence labeling baseline method are also shown

in Table 6 for comparison. Since density descriptor and modality characteristics occur very infrequently in the dataset (5 and 2 times, respectively), we do not report the results for these two FEs. For extracting SFEs associated with triggers, we see that $\text{Query}_{\text{find + desc}}$ helps in achieving a better performance than $\text{Query}_{\text{find}}$ for all elements (except for Hedge) in the case of MIMIC+Cased model. Whereas, for Uncased and Cased variants, $\text{Query}_{\text{find + desc}}$ performed better for some of such SFEs. We also note that the performance of less frequent FEs: Reason and Associated Process improved to 49.81 and 54.63 compared to baseline system's F1 (0 for both). For the majority of the FEs associated with a radiological entity, the average F1 scores lie in the range of 60-75. However, for Laterality and Size Descriptor, the values are relatively high with the highest F1 scores being 89.35 and 78.98, respectively.

The results for target entity extraction are shown in Table 7. We observe that the best F1 score for identifying the spatial triggers obtained by our proposed method (90.07) is around 12 points higher compared to the baseline system's performance of 77.89 (as reported before [4]). A later work [29] improved on the sequence labeling baseline specifically for extracting the spatial triggers (and not the FEs) where a hybrid technique combining a BERT classifier with domain constraints was employed. This improved the average F1 to 81.10 when compared with the baseline's F1 of 77.89. Thus, we see that our QA approach proposed here still outperforms the hybrid system's result by 8.9 points, though in theory a similar hybrid technique could potentially improve upon our current result further. The entity labels of the target entities are included during the FE extraction to make the queries more informative and are not part of the FE performance evaluation.

4. Discussion

The results in Tables 6 and 7 demonstrate the performance improvement in extracting spatial information from radiology reports when the problem is framed as MRC compared to traditional sequence labeling. The improvements are high: for example, improvement of average F1 scores from 65.12 to 78.13 and 71.51 to 83.77 for common FEs like Figure and Ground, respectively. This highlights the advantages of framing IE as MRC that we described in Section 1. We also note that casting IE problems as MRC is still under-explored on clinical domain datasets except for Banerjee et al. [20]. This is the only other case we are aware of MRC being used for IE from clinical reports, and there it is used only for entity extraction, not relation extraction and not with a two-turn QA approach. Moreover, we emphasize that our work covers more detailed radiological information from spatial and descriptor perspectives and extracts information from reports of multiple imaging modalities and anatomies (as opposed to previous work [1-3] focusing on either single modality or anatomy). This is the first study to use MRC both for spatial IE and for extracting important radiology information.

Our investigation using two query variations for FE extraction suggests that incorporating more information about the element in the query helps in obtaining better results, especially in cases where the meaning of FE is not obvious solely based on the FE type name. For example, we see performance improvement for Position Status for all model variants when the following description about Position Status is included in the query:

Position status refers to any position-related information, usually in context to a device. Examples include terminates and expected position.

This provides more prior knowledge about what is meant by Position Status in a radiology report context. We also find that our proposed approach tends to perform better than sequence labeling for less frequent FEs (e.g., Reason). Note that the MIMIC pre-trained BERT model underperforms both the original Uncased and Cased models for the majority of the infrequent FEs such as Associated Process, Composition Descriptor, and Size/ Measurement.

Alongside starting the queries with 'Find all', we explored two other query variations – beginning the queries with 'Get all' and 'What' (inspired by previous work [20]). Although these variations performed better than the baseline, we did not find any clear performance trend when compared to the 'Find all' variant. An exhaustive comparison of query variations could be investigated further, but that was not the focus of this work.

The moderate performance values as well as the performance variation for some FEs could be due to the infrequency of annotations in the dataset. This indicates that there is still scope for improving the results and we aim to evaluate our approach on an enlarged dataset that will have more such FEs. Although we apply our proposed method on a dataset that covers three types of radiology reports, we further intend to evaluate the generalizability of this method on multi-institutional datasets and on other imaging modalities (e.g., ultrasound and CT reports of different body parts) in a subsequent work.

5. Conclusion

We frame the problem of fine-grained radiology spatial IE as two-turn QA. This approach outperforms traditional transformer-based sequence labeling in extracting both spatial triggers and their corresponding SFEs from the radiology reports. The average F1 score for identifying spatial triggers is 90.07 and the average F1s for identifying important FEs like Figure and Ground are 78.13 and 83.77, respectively. Extracting radiology findings/devices with enough contextual information facilitates various downstream clinical applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB: R21EB029575) and the Patient-Centered Outcomes Research Institute (PCORI: ME-2018C1-10963).

References

[1]. Syeda-Mahmood T, P.D, Wong KCL, P. D, Wu JT, M. D., M. P. H, Jadhav A, P. D, Boyko O, M. D. P. D, Extracting and Learning Fine-Grained Labels from Chest Radiographs, arXiv:2011.09517 [cs]arXiv:2011.09517. URL http://arxiv.org/abs/2011.09517

[2]. Sugimoto K, Takeda T, Oh J-H, Wada S, Konishi S, Yamahata A, Manabe S, Tomiyama N, Matsunaga T, Nakanishi K, Matsumura Y, Extracting clinical terms from radiology

reports with deep learning, Journal of Biomedical Informatics 116 (2021) 103729. doi:10.1016/j.jbi.2021.103729. [PubMed: 33711545]

[3]. Steinkamp JM, Chambers C, Lalevic D, Zafar HM, Cook TS, Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning, Journal of Digital Imaging 32 (4) (2019) 554–564. doi:10.1007/s10278-019-00234-y. [PubMed: 31218554]

[4]. Datta S, Ulinski M, Godfrey-Stovall J, Khanpara S, Riascos-Castaneda RF, Roberts K, Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 2251–2260. URL https://www.aclweb.org/anthology/2020.lrec-1.274

[5]. Bradshaw T, Weisman A, Perlman S, Cho S, Automatic image classification using labels from radiology text reports: Predicting Deauville scores, Journal of Nuclear Medicine 61 (supplement 1) (2020) 1410–1410. URL https://jnm.snmjournals.org/content/61/supplement_1/1410

[6]. Wood D, Guilhem E, Montvila A, Varsavsky T, Kiik M, Siddiqui J, Kafiabadi S, Gadapa N, Busaidi AA, Townend M, Patel K, Barker G, Ourselin S, Lynch J, Cole J, Booth T, Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM), in: Medical Imaging with Deep Learning, 2020. URL https://openreview.net/forum?id=9exoP7PDD3

[7]. Wheater E, Mair G, Sudlow C, Alex B, Grover C, Whiteley W, A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records, BMC Medical Informatics and Decision Making 19 (1) (2019) 184. doi:10.1186/s12911-019-0908-7. [PubMed: 31500613]

[8]. Rubin DL, Willrett D, O'Connor MJ, Hage C, Kurtz C, Moreira DA, Automated Tracking of Quantitative Assessments of Tumor Burden in Clinical Trials, Translational Oncology 7 (1) (2014) 23–35. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998692/ [PubMed: 24772204]

[9]. Yan K, Wang X, Lu L, Summers RM, DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning, Journal of Medical Imaging 5 (3). doi:10.1117/1.JMI.5.3.036501.

[10]. Hassanpour S, Langlotz CP, Information extraction from multi-institutional radiology reports, Artificial intelligence in medicine 66 (2016) 29–39. doi:10.1016/j.artmed.2015.09.007. [PubMed: 26481140]

[11]. Bozkurt S, Alkim E, Banerjee I, Rubin DL, Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm, Journal of Digital Imaging 32 (4) (2019) 544–553. doi:10.1007/s10278-019-00237-9. [PubMed: 31222557]

[12]. Xu J, Li Z, Wei Q, Wu Y, Xiang Y, Lee H-J, Zhang Y, Wu S, Xu H, Applying a deep learning-based sequence labeling approach to detect attributes of medical concepts in clinical text, BMC Medical Informatics and Decision Making 19 (5) (2019) 236. doi:10.1186/s12911-019-0937-2. [PubMed: 31801529]

[13]. Levy O, Seo M, Choi E, Zettlemoyer L, Zero-Shot Relation Extraction via Reading Comprehension, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), 2017, pp. 333–342. doi:10.18653/v1/K17-1034.

[14]. Li X, Feng J, Meng Y, Han Q, Wu F, Li J, A Unified MRC Framework for Named Entity Recognition, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5849–5859. doi:10.18653/v1/2020.acl-main.519.

[15]. Liu J, Chen Y, Liu K, Bi W, Liu X, Event Extraction as Machine Reading Comprehension, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1641–1651. doi:10.18653/v1/2020.emnlp-main.128.

[16]. Li X, Yin F, Sun Z, Li X, Yuan A, Chai D, Zhou M, Li J, Entity-Relation Extraction as Multi-Turn Question Answering, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1340–1350. doi:10.18653/v1/P19-1129.

[17]. Li F, Peng W, Chen Y, Wang Q, Pan L, Lyu Y, Zhu Y, Event Extraction as Multi-turn Question Answering, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 829–838. doi:10.18653/v1/2020.findings-emnlp.73.

[18]. Wang XD, Weber L, Leser U, Biomedical Event Extraction as Multi-turn Question Answering, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, 2020, pp. 88–96. doi:10.18653/v1/2020.louhi-1.10.

[19]. Sun C, Yang Z, Wang L, Zhang Y, Lin H, Wang J, Biomedical named entity recognition using BERT in the machine reading comprehension framework, arXiv:2009.01560 [cs]arXiv:2009.01560. URL http://arxiv.org/abs/2009.01560

[20]. Banerjee P, Pal KK, Devarakonda M, Baral C, Knowledge Guided Named Entity Recognition for BioMedical Text, arXiv:1911.03869 [cs]arXiv:1911.03869. URL http://arxiv.org/abs/1911.03869

[21]. Bastianelli E, Croce D, Basili R, Nardi D, UNITOR-HMM-TK: Structured Kernel-based learning for Spatial Role Labeling, in: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, pp. 573–579. URL https://www.aclweb.org/anthology/S13-2096

[22]. Devlin J, Chang M-W, Lee K, Toutanova K, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[23]. Johnson AE, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016) 160035. doi:10.1038/sdata.2016.35. [PubMed: 27219127]

[24]. Langlotz CP, RadLex: A new method for indexing online educational materials, Radiographics 26 (6) (2006) 1595–1597. doi:10.1148/rg.266065168. [PubMed: 17102038]

[25]. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv:1907.11692 [cs]arXiv:1907.11692. URL http://arxiv.org/abs/1907.11692

[26]. Qu C, Yang L, Qiu M, Croft WB, Zhang Y, Iyyer M, BERT with History Answer Embedding for Conversational Question Answering, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1133–1136. doi:10.1145/3331184.3331341.

[27]. Wang Z, Ng P, Ma X, Nallapati R, Xiang B, Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering, EMNLP/IJCNLPdoi:10.18653/v1/D19-1599.

[28]. Si Y, Wang J, Xu H, Roberts K, Enhancing clinical concept extraction with contextual embeddings, Journal of the American Medical Informatics Association (2019) 1–8 doi:10.1093/jamia/ocz096. [PubMed: 30590540]

[29]. Datta S, Roberts K, A Hybrid Deep Learning Approach for Spatial Trigger Extraction from Radiology Reports, Proceedings of the Conference on Empirical Methods in Natural Language Processing 2020 (2020) 50–55. doi:10.18653/v1/2020.splu-1.6. [PubMed: 33336212]

Summary Table

Already known:

- Formulating IE tasks as question answering helps obtain improved results over traditional methods such as sequence labeling in the general and biomedical literature domains though its impact in the clinical domain is less clear.
- Important clinical information are extracted from radiology reports using NLP for different informatics applications.

What this study added:

- We proposed a two-turn QA framework based on BERT to extract important spatial and descriptor information from radiology reports by answering questions using report text. This improved the best-reported results (using a BERT-based sequence tagger) on the spatial IE task.
- This is the first study to apply a multi-turn QA approach for extracting spatial information (spatial triggers and spatial frame elements associated with both triggers and radiological entities) as well as for identifying granular information from radiology reports.
- We extracted more detailed contextual information from the reports including clinical findings, medical devices, anatomical structures, potential diagnoses, various finding and anatomy-specific radiology descriptors, as well as uncertainty and negation phrases associated with each identified finding and diagnosis.

- Two-turn question answering for extracting spatial relations from radiology reports
- First turn extracts radiology entities (e.g., finding, device) and spatial triggers
- Second turn extracts spatial relations from these terms
- Incorporating domain knowledge in the question improves performance
- Better performance compared to a standard BERT-based baseline

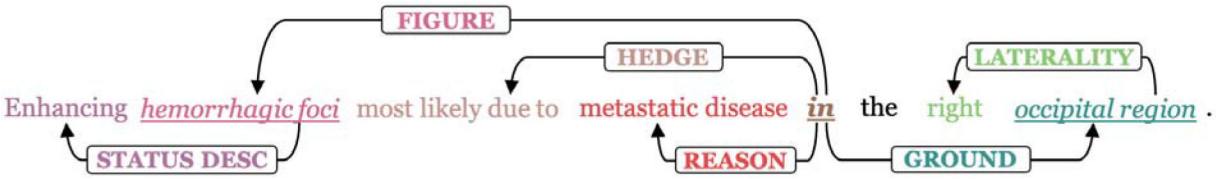
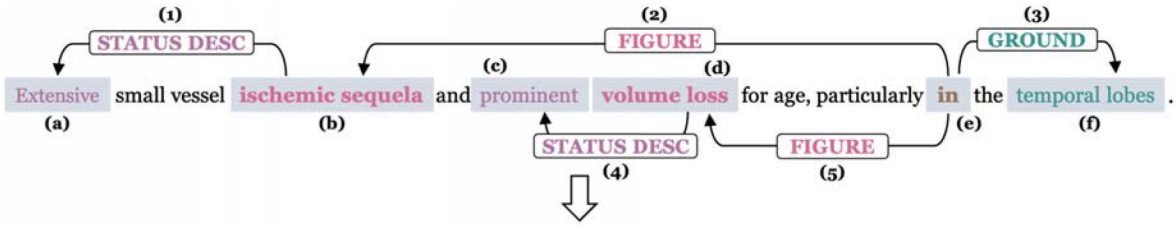


Figure 1:

Spatial and descriptive frame elements in radiology text. Figure, Ground, Hedge, and Reason are the spatial frame elements of the frame instantiated by the spatial trigger *in*. *Enhancing* denotes a status descriptor and is a descriptive element of the frame evoked by the finding entity *hemorrhagic foci* whereas *right* is the laterality and is a spatial frame element of the frame evoked by the anatomical entity *occipital region*. The *underlined and italicized* texts indicate the lexical units of the frames.



		Queries	Answer spans
Turn 1 Entities	b, d	find all clinical finding entities in the context	*ischemic sequela, volume loss*
	e	find all spatial trigger entities in the context	*in*
	f	find all anatomical structure entities in the context	*temporal lobes*
	a, c	find all descriptor entities in the context	*extensive, prominent*
	-	find all medical device entities in the context	NONE
Turn 2 Frame Elements	2, 5	find all clinical finding entities that have a figure relationship with spatial trigger entity ***in***	*ischemic sequela, volume loss*
	3	find all anatomical structure entities that have a ground relationship with spatial trigger entity ***in***	*temporal lobes*
	1	find all descriptor entities that have a status descriptor relationship with clinical finding entity ***ischemic sequela***	*extensive*
	4	find all descriptor entities that have a status descriptor relationship with clinical finding entity ***volume loss***	*prominent*
	-	find all medical device entities that have a figure relationship with spatial trigger entity ***in***	NONE

Figure 2:

Overview of two-turn QA approach for radiology spatial information extraction. Entities a-f are extracted in turn 1 and frame elements 1-5 are extracted in turn 2. The ***bold and italicized*** texts in the queries for turn 2 indicate that they are extracted from turn 1. Only a subset of queries for which there is no answer (indicated using NONE) are shown for brevity. The example sentence contains two radiographic findings–*ischemic sequela* and *volume loss*, described through *extensive* and *prominent*, respectively.

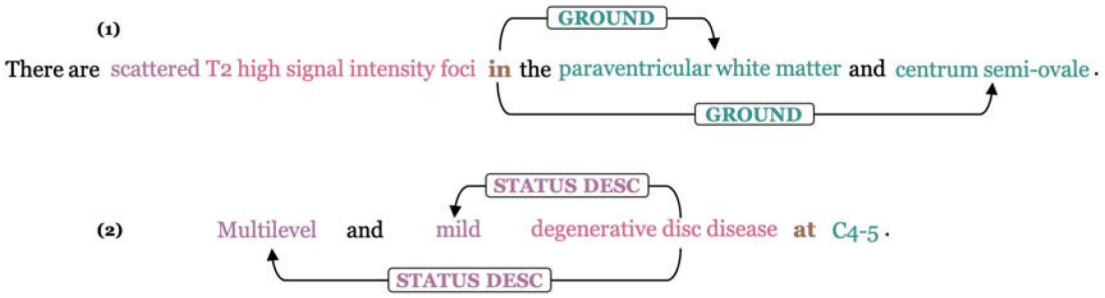


Figure 3:

(1) Two Ground elements are linked to a spatial trigger and (2) two status descriptors are linked to a radiological finding. For (1), the query for extracting anatomical locations with respect to the spatial trigger *in* should return two spans – *paraventricular white matter* and *centrum semi-ovale*. For (2), an MRC model is expected to return the spans corresponding to *Multilevel* and *mild* as output when queried for the status descriptive elements of *degenerative disc disease*.

Table 1:

Target entities extracted in turn 1.

Target entity type	Description	RadLex Class
Spatial trigger	Spatial prepositions (e.g., *in*), verbs (e.g., *demonstrate*), verb followed by prepositions (e.g., *projected at*), etc.	Not applicable
Finding	Terms related to radiological observations, clinical findings (including those suggesting diagnoses)	Clinical finding, Imaging observation
Anatomy	Anatomical location	Anatomical entity
Device	Medical device	Medical device
Tip	Tip of a medical device	Portion of medical device
Location descriptor	Describing how a finding is located with respect to an anatomy	Location descriptor
Other descriptor	Modifiers describing a radiological observation or finding	RadLex descriptor (except Location and Certainty)
Assertion	Uncertainty and negated phrases used by radiologists	Certainty descriptor
Position	Position status of a device (e.g., *good position*)	Not applicable
Quantity	Any quantitative term in the report text (e.g., *3 mm*)	Not applicable
Process	Describing motion, change, etc.	Process

Table 2:

Frame Elements extracted in turn 2, their descriptions, and associated entity types. ST - Spatial Trigger. Desc - Descriptor.

Frame Elements		Description	Entity Types of Related Entities
SPATIAL	Figure	Object whose location is described	ST; Finding/Anatomy/Device/Tip
	Ground	Anatomical location of Figure	ST; Anatomy
	Hedge	Uncertainty expressions used by radiologists	ST; Assertion
	Diagnosis	Clinical condition or disease associated with a radiological finding	ST; Finding
	Position Status	Any position-related information, usually in context to a device	ST; Position
	Relative Position	Terms used for describing the orientation of a radiological entity wrt to an anatomical location	ST; Location descriptor
	Distance	Actual distance of finding or device from the anatomical location	ST; Quantity
	Reason	Clinical condition or disease that acts as the source of a radiological finding	ST; Finding
	Associated Process	Any process or activity associated with a spatial relation	ST; Process
	Morphologic	Indicates shape	Finding/Anatomy; Desc
	Size Desc	Indicates size description	Finding/Anatomy/Device; Desc
	Distribution Pattern	Indicates distribution patterns	Finding/Anatomy; Desc
	Composition	Indicates composition of a radiological finding	Finding/Anatomy; Desc
	Laterality	Indicates side	Finding/Anatomy/Device; Desc
	Size/Measurement	Actual size of a finding	Finding; Desc
DESC	Status	Indicates status of entities	Finding/Anatomy/Device; Desc
	Quantity	Indicates quantity of a radiological entity	Finding/Anatomy/Device; Desc
	Temporal	Indicates temporality	Finding/Device; Desc
	Negation	The associated negated phrase	Finding/Anatomy; Desc

Table 3:

Modified entity types to be used in queries.

Target entity type	Modified entity type
Finding	Clinical finding
Anatomy	Anatomical structure
Device	Medical device
Tip	Medical device tip
Other descriptor	Descriptor
Assertion	Assertion-related
Position	Position-related
Process	Associated process

Table 4:

Query template and example. Q_f: $\text{Query}_{\text{find}}$. Q_{f+d}: $\text{Query}_{\text{find + desc}}$.

Extraction step	Query template	Example
Target entities	***Entity type* :** ENT Q_f**:** find all **ENT** entities in the context.	***Entity type* :** Spatial trigger Q_f **:** find all **spatial trigger** entities in the context.
Spatial and descriptive frame elements	***Frame element type* :** REL ***Target Entity type* :** ENT_1 **ENT_1 span from turn 1 :** ENT_1_SPAN ***Related Entity type* :** ENT_2 Q_f **:** find all **ENT_2** entities in the context that have a/an **REL** relationship with **ENT_1** entity **ENT_1_SPAN**. Q_{f+d} **:** a general **description** about **REL +** Q_f	***Frame element type* :** Figure ***Target Entity type* :** Spatial trigger **Spatial trigger span from turn 1 :** *in* ***Related Entity type* :** Clinical finding Q_f **:** find all **clinical finding** entities in the context that have a **figure** relationship with **spatial trigger** entity ***in***. Q_{f+d} **: Figure** refers to finding or device or tip entities that are described with respect to an anatomical structure. **+** Q_f

Table 5:

Hyperparameters used in the experiments.

Parameter	Value
Sliding window size for context passage	200
Overlap between adjacent windows	45
Maximum number of training epochs	10
Learning rate	2e–5
Trade-off between two turns	0.25
Maximum norm for gradients	1
Warmup ratio	0.1

Table 6:

Average F1 measures of $BERT_{LARGE}$ models over 10-fold CV for spatial and descriptive frame element extraction. DESC: Descriptive. Q_f: $Query_{find}$. Q_{f+d}: $Query_{find + desc}$. M+Cased: MIMIC+Cased. Count: Number of annotations in the dataset. Dash (−): not available for baseline method.

FRAME ELEMENTS		Proposed approach						Baseline M+Cased	Count
		Uncased		Cased		M+Cased			
		Q_f	Q_{f+d}	Q_f	Q_{f+d}	Q_f	Q_{f+d}		
SPATIAL	Figure	**78.13**	77.29	76.72	77.57	76.44	77.40	65.12	1491
	Ground	83.76	83.40	83.31	82.27	83.17	**83.77**	71.51	1537
	Hedge	75.47	76.44	**77.18**	76.42	75.90	74.97	57.82	249
	Diagnosis	69.32	73.32	**73.94**	72.67	65.47	67.92	50.76	190
	Position Status	68.72	68.75	66.98	67.12	68.43	**70.37**	60.37	167
	Relative Position	77.19	76.42	**77.53**	76.71	75.78	76.15	66.33	398
	Distance	84.65	86.54	85.36	85.20	87.94	**90.09**	88.05	45
	Reason	39.51	32.34	39.51	**49.81**	17.71	44.89	0	33
	Associated Process	48.52	**54.63**	43.15	42.29	38.95	41.36	0	21
	Morphologic	52.48	58.14	49.92	**60.52**	48.04	45.53	–	69
	Size Desc	76.16	73.80	78.16	78.94	78.56	**78.98**	–	93
	Distribution Pattern	57.45	63.62	59.74	64.01	59.22	**66.03**	–	65
	Composition	41.46	33.63	41.67	**46.88**	26.49	20.48	–	17
	Laterality	88.43	88.51	**89.35**	87.49	87.78	87.32	–	612
	Size/Measurement	45.43	**48.44**	41.51	43.59	34.46	32.06	–	23
DESC	Status	**64.67**	62.60	63.38	61.67	59.17	59.09	–	452
	Quantity	72.56	72.32	72.82	71.61	72.47	**73.11**	–	130
	Temporal	70.87	70.63	70.5	71.47	67.31	**71.78**	–	113
	Negation	58.08	61.06	**67.75**	65.04	60.95	61.83	–	103

Table 7:

Average F1 measures of $BERT_{LARGE}$ models over two 10-fold CVs for target entity extraction. M+Cased: MIMIC+Cased.

Target entities	Uncased	Cased	M+Cased
Spatial trigger	89.99	89.50	**90.07**
Finding	76.89	**78.26**	76.11
Anatomy	**87.56**	87.40	87.46
Device	91.87	92.68	**93.12**
Tip	99.18	98.41	**99.32**
Location descriptor	**81.50**	81.21	80.89
Other descriptor	84.19	**84.24**	84.09
Assertion	78.48	**80.85**	79.40
Position	69.68	71.41	**72.81**
Quantity	**85.54**	85.37	83.23
Process	**60.93**	59.26	60.19