Data in Brief 32 (2020) 106056

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A dataset of chest X-ray reports annotated with Spatial Role Labeling annotations

Surabhi Datta, Kirk Roberts*

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

ARTICLE INFO

Article history:
Received 23 June 2020
Revised 14 July 2020
Accepted 16 July 2020
Available online 25 July 2020

Keywords:
Spatial relations
Spatial Role Labeling
Radiology report
Chest radiology
Natural language processing
Information extraction

ABSTRACT

In this paper, we present a dataset consisting of 2000 chest X-ray reports (available as part of the Open-i image search platform) annotated with spatial information. The annotation is based on Spatial Role Labeling. The information includes annotating a radiographic finding, its associated anatomical location, any potential diagnosis described in connection to the spatial relation (between finding and location), and any hedging phrase used to describe the certainty level of a finding/diagnosis. All these annotations are identified with reference to a spatial expression (or SPATIAL INDICATOR) that triggers a spatial relation in a sentence. The spatial roles used to encode the spatial information are TRAJECTOR, LANDMARK, DIAGNOSIS, and HEDGE. In total, there are 1962 SPATIAL INDICATORS (mainly prepositions). There are 2293 TRAJECTORS, 2167 LANDMARKS, 455 DIAGNOSIS, and 388 HEDGES in the dataset. This annotated dataset can be used for developing automatic approaches targeted toward spatial information extraction from radiology reports which then can be applied to numerous clinical applications. We utilize this dataset to develop deep learning-based methods for automatically extracting the SPATIAL INDICATORS as well as the associated spatial roles [1].

DOI of original article: 10.1016/j.jbi.2020.103473
* Corresponding author.
E-mail addresses: surabhi.datta@uth.tmc.edu (S. Datta), kirk.roberts@uth.tmc.edu (K. Roberts).

https://doi.org/10.1016/j.dib.2020.106056

Specifications Table

Subject	Health Informatics
Specific subject area	Spatial information extraction from chest X-ray reports based on Spatial Role Labeling schema for spatial language understanding in radiology reports
Type of data	Table, Figure, Text, Annotated data in XML format
How data were acquired	A subset of 2000 chest X-ray reports were used from a pool of 3996 de-identified reports collected from the Indiana Network for Patient Care (available as one of the Open-i datasets released by the National Library of Medicine.)
Data format	Raw, Processed
Parameters for data collection	2000 chest X-ray reports that are annotated with important spatial information were selected from the set of 2470 non-normal reports in the Open-i chest X-ray report dataset as adjudicated by two annotators.
Description of data collection	These 2000 reports were annotated with four spatial roles using the Brat toolkit. First, the spatial indicators (usually the spatial prepositions) triggering any spatial relation between a radiographic finding and an anatomical location were annotated for each sentence. Then, four spatial roles–the radiographic finding, its corresponding location, hedging phrase, and any potential diganosis were annotated with respect to a specific spatial indicator.
Data source location	Primary data source: Open-i chest X-ray dataset (https://openi.nlm.nih.gov/). *Associated research paper: "Preparing a collection of radiology examinations for distribution and retrieval"*–https://doi.org/10.1093/jamia/ocv080
Data accessibility	Repository name: Mendeley data repository Data identification number: 10.17632/yhb26hfz8n.1 Direct URL to data: https://doi.org/10.17632/yhb26hfz8n.1, https://github.com/krobertslab/datasets/tree/master/rad-sprl
Related research article	S. Datta, Y. Si, L. Rodriguez, S. E. Shooshan, D. Demner-Fushman, K. Roberts, Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning, Journal of Biomedical Informatics 108 (2020) 103473. doi:10.1016/j.jbi.2020.103473.

Value of the Data

- The spatial information annotated in this dataset captures clinically significant information of chest X-ray imaging results. This annotation schema proposes a way to encode radiological spatial knowledge from report text. The annotated information includes the main radiographic finding detected, the anatomical location where the finding has been described to be present, any diagnosis associated with the finding-location pair, as well as any hedging phrase used to suggest the diagnosis or the finding.
- The dataset can be used to develop automatic NLP systems for extracting spatial information from radiology reports. These systems have the potential to facilitate various clinical applications. A few of these include easy visualization of contextual information associated with abnormal radiographic findings from a spatial perspective, automatic tracking of findings, and automatic annotation of corresponding radiographic images with spatial and diagnosis information.
- The models developed on this dataset could be further leveraged by applying them on other types of radiology reports belonging to different imaging modality such as chest Computed Tomography (CT) scans and Magnetic Resonance Imaging (MRI) as the annotated information types are common across different modalities and/or anatomies.

1. Data Description

This 2000 chest X-ray reports dataset is a subset of 3996 reports collected from the Indiana Network for Patient Care [2]. Specifically, the 2000 report subset is composed from the set of 2470 non-normal reports as judged by two human annotators. The annotation schema

Table 1
Annotated dataset descriptions.

Attribute	Description
Document	Represents a chest X-ray report
Text	Raw text of the report
Annotations	Contains the processed text and spatial annotations for a report
Token	Contains start character and number of characters of a token
Sentence	Contains start token number and number of included tokens to identify a sentence
RadSpRLRelation	Indicates the presence of a spatial relation. Includes the start token number and number of tokens of a spatial expression (SPATIAL INDICATOR) in a sentence, also contains all the associated spatial roles with respect to this SPATIAL INDICATOR
Spatial roles under RadSpRLRelation	
TRAJECTOR	Radiological entity (usually a radiographic finding whose position is described
LANDMARK	Anatomical location of a TRAJECTOR
DIAGNOSIS	Potential diagnosis associated with a spatial relation
HEDGE	Any uncertainty phrase used to describe a finding or diagnosis

Table 2
Spatial indicator statistics.

Parameter	Frequency
Total number of SPATIAL INDICATORS	1962
Number of distinct SPATIAL INDICATORS	29
Most frequent indicators	
of	765
in	526
without	176
with	141
within	102

is based on Spatial Role Labeling (SpRL) [3,4] and has been extended to encode information in radiology context. This includes identifying a SPATIAL INDICATOR in a sentence and consequently annotating the main radiographic finding and anatomical location that are connected by this SPATIAL INDICATOR. Additionally, the spatial annotations include any potential diagnosis identified in a sentence with reference to the spatial relation between a finding and a location. The annotations also include any uncertainty phrase or hedge used to describe a finding/diagnosis. These four information types denote the four spatial roles with respect to a SPATIAL INDICATOR in a sentence. The schema is referred to as Rad-SpRL. The dataset is included in XML format (available at https://doi.org/10.17632/yhb26hfz8n.1 in the Mendeley data repository and https://github.com/krobertslab/datasets/) and the relevant details are described in Table 1. A few details of the SPATIAL INDICATORS in the dataset are included in Table 2. In total, there are 29 unique spatial expressions. The most frequent phrases for each of the four spatial roles annotated are shown in Table 3. We also note the frequent descriptors used in describing roles like TRAJECTOR and DIAGNOSIS. Note that '*XXXX*' is used to denote any de-itentified term in the report text. For each of DIAGNOSIS, TRAJECTOR, and LANDMARK, the most common associated other two spatial roles are demonstrated in Figs. 1–3. We provide a brief statistics on the terms that are annotated as two different spatial roles depending on the context in a sentence in Table 4. We also analyze the terms expressing HEDGE role (illustrated in Table 5).

2. Experimental Design, Materials, and Methods

In this dataset, we attempt to widen the scope of clinically significant information types to be extracted from chest X-ray reports and additionally aim to relate all the information in context

Table 3
Most frequent terms for each spatial role.

Spatial Role	Term	**Frequency** (descriptors contained in Term)
TRAJECTOR	*opacity*	279 (nodular, streaky, interstitial, focal airspace, focal, airspace, vague, patchy, bibasilar, ill-defined, mild streaky, subtle increased, few small nodular, round, scattered, rounded nodular, abnormal, vague nodular, patchy airspace, bilateral, bandlike, vague patchy, dense, minimal, minimal streaky, streaky basilar, alveolar)
	degenerative change	205 (mild, minimal, diffuse, moderate, severe, multilevel, chronic, advanced)
	pneumothorax	63 (moderate right-sided, large)
	pleural effusion	63 (large, small bilateral, large right)
	consolidation	57 (focal, focal airspace, dense)
Total Distinct		861
LANDMARK	*lung*	285 (also includes lungs)
	thoracic spine	146 (mid, lower)
	spine	111
	left lung base	43
	thorax	40
Total Distinct		570
DIAGNOSIS	*scarring*	53 (pleural, pleural-parenchymal, chronic)
	atelectasis	83 (subsegmental, focal, chronic subsegmental, foci of subsegmental, lingular)
	infiltrate	21 (focal)
	granuloma	15 (calcified, partially calcified)
	emphysema	11
Total Distinct		224
HEDGE	*may represent*	40
	XXXX	39
	consistent with	38 (focal)
	XXXX represent	34 (also includes XXXX represents, XXXX representing, XXXX representative of)
	compatible with	21
Total Distinct		80

to a spatial relation between a finding and a location. This provides more contextual information about a radiographic finding. Many of the previous works on radiology information extraction mainly focused on extracting radiological entities (findings, diagnoses, etc.) separately without establishing any relation among these entities [5,6,8,7].

We further analyze the variations of SPATIAL INDICATORS in the dataset. Besides the five most frequent ones mentioned in Table 2, the other spatial prepositions include – '*at*', '*over*', '*on*', '*throughout*', '*under*', '*along*', '*near*', '*to*', '*through*', '*between*', '*adjacent*', '*beneath*', '*from*', '*into*', '*below*', '*above*', '*around*', '*towards*', '*about*', '*behind*'. This dataset also includes four more verbal spatial expressions – '*overlie*', '*overlies*', '*overlying*', and '*involving*'. However, these four expressions occur very infrequently and together account for 30 out of 1962 SPATIAL INDICATORS. Also, note that the indicator '*without*' denotes a negated spatial relation and is oftentimes present as part of the common negated phrase used in radiology reports – '*without evidence of*'.

We inspect the dataset to analyze the most frequent terms annotated for each spatial role and observe that the top five frequent TRAJECTORS are different from the five most frequent DIAGNOSIS terms (as illustrated in Table 3). There are more distinct TRAJECTORS and LANDMARKS than DIAGNOSIS and HEDGE terms.

We also analyze, for each spatial role, the most frequently associated other roles (Figs. 1–3). For this, we consider three terms among the five most frequent terms (shown in Table 3) for each role. It is interesting to observe that no diagnoses are associated with three frequent radio-

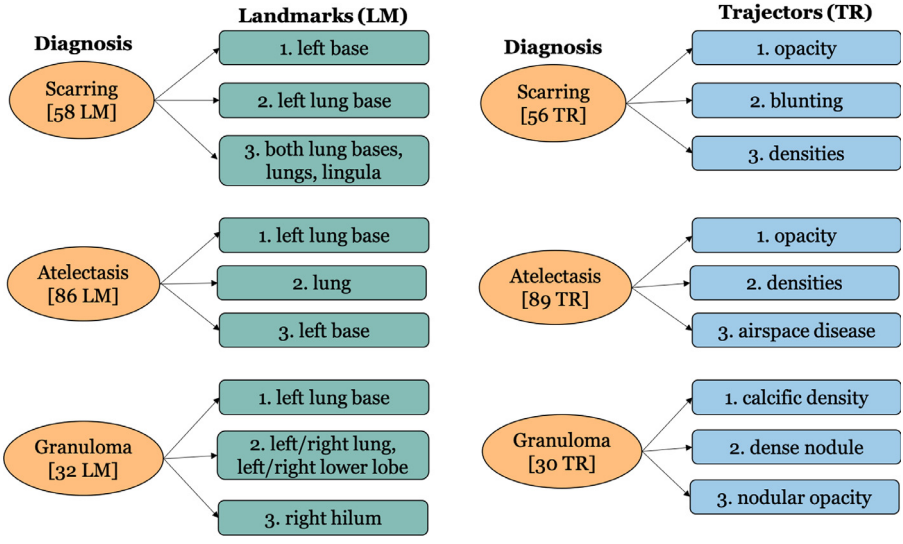

Fig. 1. Most common associated landmarks and trajectors for three frequent DIAGNOSIS. [*n* LM] indicates that a particular diagnosis is connected to a total of *n* landmarks, while [*n* TR] indicates that a particular diagnosis is connected to a total of *n* trajectors.

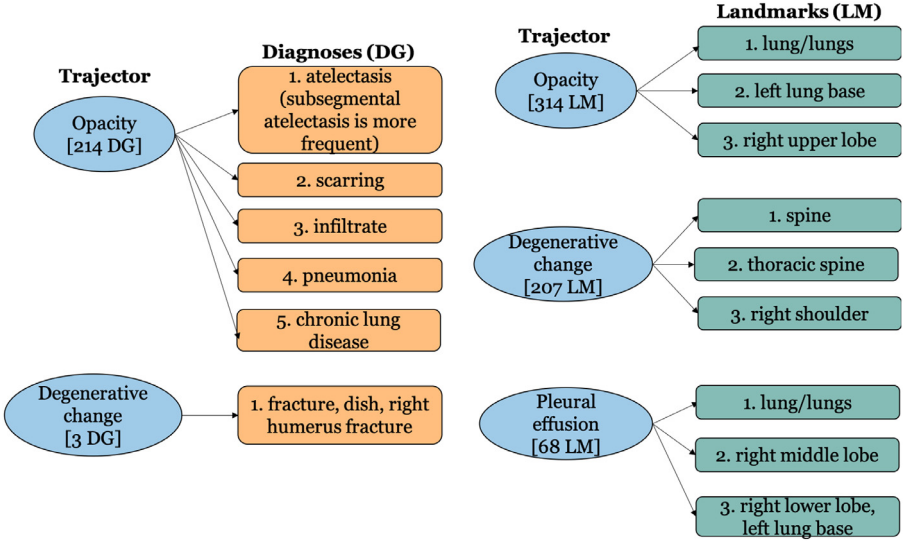

Fig. 2. Most common associated diagnoses and landmarks for three frequent TRAJECTOR. [*n* DG] indicates that a particular trajector is connected to a total of *n* dagnoses, while [*n* LM] indicates that a particular trajector is connected to a total of *n* landmarks.

graphic findings (TRAJECTORS) – '*pneumothorax*', '*pleural effusion*', and '*consolidation*' (as shown in Fig. 2).

In the process of annotating the reports, we noticed that some terms take different spatial roles depending on the context. We then inspect this overlap between two spatial roles in our annotated dataset. Specifically, the overlapping characteristics between TRAJECTOR and DIAGNOSIS as well as between TRAJECTOR and LANDMARK are shown in Table 4. There are more distinct

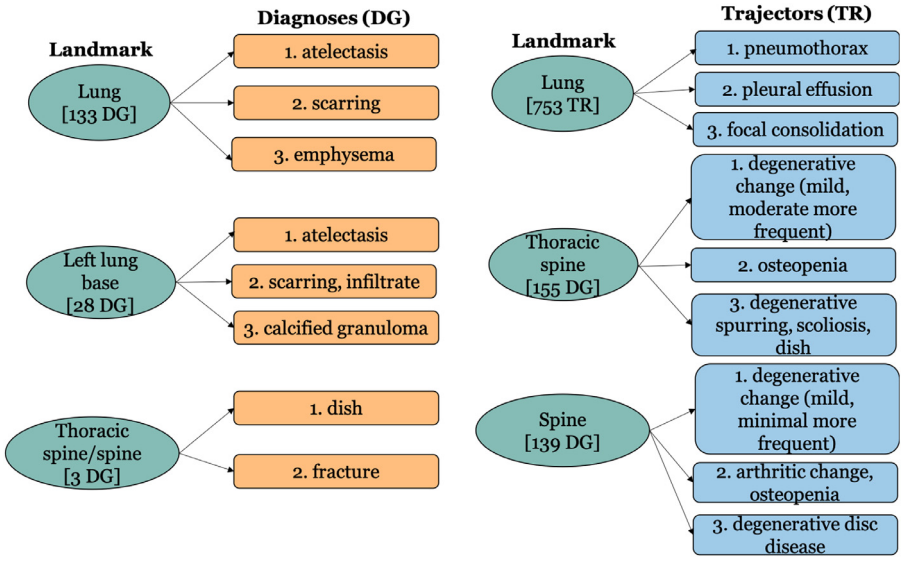

Fig. 3. Most common associated diagnoses and trajectors for three frequent LANDMARK. [*n* DG] indicates that a particular landmark is connected to a total of *n* dagnoses, while [*n* TR] indicates that a particular landmark is connected to a total of *n* trajectors.

terms that have overlap between TRAJECTOR and LANDMARK than between TRAJECTOR and DIAGNOSIS. Around 52% of the terms that act as both TRAJECTOR and LANDMARK an equal number of times oftentimes have the same text span and are related to anatomical structures or portions. Consider the following example:

Visualized ***osseous structures****of the thorax are without acute abnormality.*

Here, '*osseous structures*' act as both TRAJECTOR and LANDMARK. It takes the role of a TRAJECTOR when considered in relation to the indicator '*of*' and acts as a LANDMARK when considered in relation to '*without*'.

Additionally, we note that the terms that are annotated as both TRAJECTOR and LANDMARK appear more often as a LANDMARK than a TRAJECTOR (as shown in Table 4). There are certain findings like '*pleural thickening*/*thickening*' which appear both as TRAJECTOR and DIAGNOSIS with the same frequency.

Since the hedging terms are used both in context to describing a radiographic finding as well as a diagnosis, we intend to investigate their distribution in both the cases. We find that certain phrases such as '*probable*' and '*or*' are more representative of describing the findings rather than diagnoses. We also witness a variety of hedging expressions that occur rarely in the dataset. Besides the ones presented in Table 5, few other rare hedging phrases include – '*possibly related to*', '*is a consideration*', '*favored as*', '*could be secondary to*', and '*cannot be ruled out*'.

Table 4
Overlapping terms between two spatial roles.

Parameter	Frequency
Distinct overlapping terms (TRAJECTOR and DIAGNOSIS)	45
Distinct overlapping terms (TRAJECTOR and LANDMARK)	73
Same terms with equal frequency (TRAJECTOR and LANDMARK)	38
Terms appearing more as TRAJECTORand less as DIAGNOSIS	
Term	**Frequency**
infiltrate/focal infiltrate	TRAJECTOR:40 DIAGNOSIS:19
calcified granuloma/calcified granulomas	TRAJECTOR:37 DIAGNOSIS:6
focal airspace disease	TRAJECTOR:32 DIAGNOSIS:2
bronchovascular crowding	TRAJECTOR:22 DIAGNOSIS:2
fracture/fractures	TRAJECTOR:18 DIAGNOSIS:3
nodule	TRAJECTOR:18 DIAGNOSIS:2
Terms appearing more as DIAGNOSISand less as TRAJECTOR	
Term	**Frequency**
scarring	DIAGNOSIS:44 TRAJECTOR:21
atelectasis	DIAGNOSIS:43 TRAJECTOR:14
subsegmental atelectasis	DIAGNOSIS:23 TRAJECTOR:7
emphysema	DIAGNOSIS:9 TRAJECTOR:4
Terms appearing more as LANDMARKand less as TRAJECTOR	
Term	**Frequency**
right upper lobe	LANDMARK:35 TRAJECTOR:2
right	LANDMARK:27 TRAJECTOR:2
right base	LANDMARK:15 TRAJECTOR:2
Common terms appearing both as TRAJECTORand LANDMARKwith equal frequency	
Term	**Frequency**
osseous structures	31
region	5
peripheral aspect	4

Table 5
Analysis of HEDGE terms.

Description	Terms
Frequent HEDGES that appear without DIAGNOSIS	*possible/possibly, or, probable/probably, appears to be, versus*
Frequent HEDGES that appear with DIAGNOSIS	*may represent, consistent with, XXXX, compatible with, XXXX representing*
HEDGES that only appear when there is no DIAGNOSIS	*apparent, questionable, and/or, probable, suggestion of, difficult to exclude, or XXXX, or, approximately, apparently, cannot be excluded* (11)
Example Hedges that only appear once	*may be partially due to, favored to represent, cannot be excluded, raise concern for, difficult to exclude*

3. Ethics statement

This work includes chest X-ray reports of patients collected from the Indiana Network for Patient Care in a previous study [2]. The reports are de-identified and do not involve experimentation with human subjects.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB: R21EB029575), the U.S. National Library of Medicine (NLM: R00LM012104), the Patient-Centered Outcomes Research Institute (PCORI: ME-2018C1-10963) and the Cancer Prevention Research Institute of Texas (CPRIT: RP160015).

References

[1] S. Datta, Y. Si, L. Rodriguez, S.E. Shooshan, D. Demner-Fushman, K. Roberts, Understanding spatial language in radiology: representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning, J. Biomed. Inform. 108 (2020) 103473, doi:10.1016/j.jbi.2020.103473.

[2] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inf. Assoc. 23 (2) (2016) 304–310, doi:10.1093/jamia/ocv080.

[3] P. Kordjamshidi, M.V. Otterlo, M.-F. Moens, Spatial role labeling: task definition and annotation scheme, in: Proceedings of the Language Resources & Evaluation Conference, 2010, pp. 413–420.

[4] P. Kordjamshidi, T. Rahgooy, U. Manzoor, Spatial language understanding with multimodal graphs using declarative learning based programming, in: Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing, 2017, pp. 33–43, doi:10.18653/v1/w17-4306.

[5] S. Hassanpour, G. Bay, C.P. Langlotz, Characterization of change and significance for clinical findings in radiology reports through natural language processing, J. Digit. Imaging 30 (3) (2017) 314–322, doi:10.1007/s10278-016-9931-8.

[6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3462–3471, doi:10.1109/CVPR.2017.369.

[7] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, PadChest: A Large Chest X-ray Image Dataset With Multi-Label Annotated Reports (2019).

[8] M. Annarumma, S.J. Withey, R.J. Bakewell, E. Pesce, V. Goh, G. Montana, Automated triaging of adult chest radiographs with deep artificial neural networks, Radiology 291 (1) (2019) 196–202, doi:10.1148/radiol.2018180921.