Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3686–3695
Marseille, 20-25 June 2022

A Cross-document Coreference Dataset for Longitudinal Tracking across Radiology Reports

Surabhi Datta, Hio Cheng Lam, Atieh Pajouhi, Sunitha Mogalla, Kirk Roberts
School of Biomedical Informatics, The University of Texas Health Science Center at Houston
{surabhi.datta,kirk.roberts}@uth.tmc.edu

Abstract

This paper proposes a new cross-document coreference resolution (CDCR) dataset for identifying co-referring radiological findings and medical devices across a patient's radiology reports. Our annotated corpus contains 5872 mentions (findings and devices) spanning 638 MIMIC-III radiology reports across 60 patients, covering multiple imaging modalities and anatomies. There are a total of 2292 mention chains. We describe the annotation process in detail, highlighting the complexities involved in creating a sizable and realistic dataset for radiology CDCR. We apply two baseline methods–string matching and transformer language models (BERT)–to identify cross-report coreferences. Our results indicate the requirement of further model development targeting better understanding of domain language and context to address this challenging and unexplored task. This dataset can serve as a resource to develop more advanced natural language processing CDCR methods in the future. This is one of the first attempts focusing on CDCR in the clinical domain and holds potential in benefiting physicians and clinical research through long-term tracking of radiology findings.

Keywords: tracking, cross-document coreference resolution, radiology

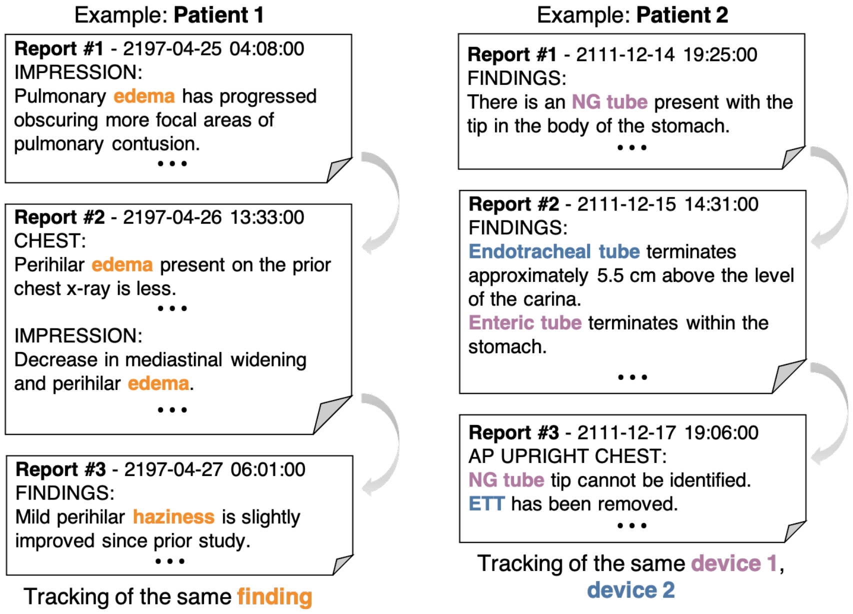
1. Introduction

Radiology reports contain rich descriptions of clinically important findings and medical devices. Oftentimes, these findings and devices are referred to multiple times in a single report and are also referred to across different reports of a patient. Radiologists make such references in multiple reports mainly to highlight any longitudinal changes of a particular finding (e.g., change in a *tumor* at a certain location) and also to describe any interval changes in a device position (e.g., change in the position of an *endotracheal tube* inserted in a patient with respect to an anatomical location). Although extracting important information (e.g., findings, anatomical locations) from radiology reports has been widely studied (Hassanpour and Langlotz, 2016; Steinkamp et al., 2019; Datta et al., 2020; Syeda-Mahmood et al., 2020; Sugimoto et al., 2021), tracking (or identifying the coreferences) of radiological findings across reports is unexplored. Automated tracking of findings and devices across a patient's radiology reports holds potential to reduce physician burden in making patient-related decisions as well as to facilitate various retrospective clinical research studies.

Tracking the same finding or device across reports is a challenging problem as it relies on radiology domain knowledge and requires understanding the linguistic variations used by radiologists as well as understanding both linguistic and domain-specific context across different reports. This is illustrated through an example in Figure 1, where, for Patient 1, the *perihilar edema* described in one of the subsequent reports of this patient is referencing to the *pulmonary edema* mentioned in a previous report, and is again described through a different expression, *perihilar haziness*, in a later report. Here, these three finding entities–*pulmonary edema*, *perihilar edema*, and *perihilar haziness* are describing the progress of the same finding for this patient. Similarly, for Patient 2, *NG tube* and *Enteric tube* are discussing the same device, whereas *Endotracheal tube* and *ETT* are describing the change in the status of another device. Thus, we see that there is a strong reliance on domain language knowledge and context information to identify the co-referring expressions of the same findings or devices across reports.

In this work, we introduce an annotated dataset to track the same radiological findings and medical devices across reports. We sample a total of 60 patients from the publicly available MIMIC-III clinical database (Johnson et al., 2016), with an average of 10.6 reports per patient. The reports include a variety of imaging modalities covering different human anatomies. We provide a detailed description of the annotation guideline in this paper. Our tracking dataset comprises of a total of 5872 mentions with 2292 mention chains. A chain here represents all the mentions across reports of a patient that refer to the same finding or device entity. We represent the tracking task with enough specificity to capture the clinical granularities that are critical to treatment planning. For example, a *fracture* detected at the right frontal lobe of the skull is different from a *fracture* detected at the left temporal lobe, and, therefore, these two fractures will be placed in two different mention chains. More details are explained in the annotation guideline (Sections 4.1 and 4.2). Instructions to access the annotated dataset are available at GitHub[1]. We employ two baseline methods–a rule-based system and a transformer

[1] `https://github.com/krobertslab/datasets/tree/master/rad-tracking`



Example: **Patient 1**

Report #1 - 2197-04-25 04:08:00
IMPRESSION:
Pulmonary **edema** has progressed obscuring more focal areas of pulmonary contusion.
• • •

Report #2 - 2197-04-26 13:33:00
CHEST:
Perihilar **edema** present on the prior chest x-ray is less.
• • •
IMPRESSION:
Decrease in mediastinal widening and perihilar **edema**.
• • •

Report #3 - 2197-04-27 06:01:00
FINDINGS:
Mild perihilar **haziness** is slightly improved since prior study.
• • •

Tracking of the same **finding**

Example: **Patient 2**

Report #1 - 2111-12-14 19:25:00
FINDINGS:
There is an **NG tube** present with the tip in the body of the stomach.
• • •

Report #2 - 2111-12-15 14:31:00
FINDINGS:
Endotracheal tube terminates approximately 5.5 cm above the level of the carina.
Enteric tube terminates within the stomach.
• • •

Report #3 - 2111-12-17 19:06:00
AP UPRIGHT CHEST:
NG tube tip cannot be identified.
ETT has been removed.
• • •

Tracking of the same **device 1**, **device 2**

Figure 1: Examples of tracking the same finding (*edema*) and the same devices (*NG tube* and *Endotracheal tube*) across multiple reports.

language-based system, BERT (Devlin et al., 2019), to automatically identify the cross-report coreferences. Finally, we evaluate the performance of the systems using standard coreference metrics.

2. Related Work

Most of the prior work using radiology report text has developed NLP systems to extract important entities such as findings, diagnoses, anatomical locations, and their respective descriptor terms (Hassanpour and Langlotz, 2016), with some focusing on more comprehensive information extraction (Steinkamp et al., 2019; Sugimoto et al., 2021; Datta and Roberts, 2022). Some studies have targeted extracting information from the reports to automatically generate labels for the corresponding medical images (Syeda-Mahmood et al., 2020; Bradshaw et al., 2020; Wood et al., 2020).

In the context of automated tracking, existing research has highlighted the requirement of a tracking system to track radiological findings. Rubin et al. (2014) has extracted tumor-related quantitative assessments to facilitate automated tracking. More recently, Bozkurt et al. (2019) has focused on automatically identifying measurements and their corresponding descriptor terms from the reports with the aim to improve care delivery by tracking the same lesions across multiple patient encounters. Another study (Steinkamp et al., 2019) that extracted various important contextual information from radiology reports has also highlighted the benefits of automatic tracking. Two studies (Mabotuwana et al., 2018; Mabotuwana et al., 2019) have concentrated on automated matching of follow-up imaging recommendations from the reports using contextual information (e.g., recommended anatomy) and various other features (e.g., text-based similarity features). Interestingly, an earlier attempt (Son et al., 2004) was made where a probabilistic model was employed to correlate lung mass or lung lesion-related findings across different computed tomography documents associated with lung cancer patients. However, the study highlighted limitations such as requirement of more refined definitions for locations (a highly weighted feature in the probabilistic model) in order to handle scenarios where multiple findings are detected around the same location. In this work, we aim to track a broad range of radiological findings and devices across reports, irrespective of their imaging modality.

We formulate the tracking problem as a cross-document (CD) coreference resolution task. In the general domain, there has been recent advancements to this relatively less explored and challenging task (Barhom et al., 2019; Cattan et al., 2021a; Cattan et al., 2021b; Bugert et al., 2021; Cattan et al., 2021c). Among the most recent contributions, Cattan et al. (2021a) developed the first end-to-end CD coreference resolution model where they applied the model over predicted mentions and achieved first baseline performances on the standard ECB+ dataset. Another work by Cattan et al. (2021b) proposed more realistic principles for evaluating CD coreference models (e.g., tackling lexical ambiguities involved in real-world CD coreferences). Cattan et al. (2021c) also proposed a hierarchical CD coreference resolution task where they identify the coreference clusters and hierarchy between them. In the medical domain, Wright-Bettner et al. (2019) provided insights on the challenging aspects of this task both from model and annotator perspectives through complex illustrative examples from a colon cancer dataset. They suggested relying more on schematic rules and less on annotator intuition to annotate more realistic and consistent CD coreference relations. Their work also highlighted the difficulties associated with creating human-annotated CD gold annotations on a sizable dataset and, thereby, restricted their annotation scope (e.g., limit the CD relations to a set of three notes per patient). In this work, we take into account some of the challenges described in the previous papers and aim to create a realistic gold-annotated CD radiology dataset with more number of reports representing a patient.

Besides the two works on CD coreference resolution on a medical corpus, (Son et al., 2004; Wright-Bettner et al., 2019), a few more studies have targeted within-document task (Apostolova et al., 2012; Miller et al., 2017). Thus, CD coreference is still under-explored in the clinical domain and we tackle this challenging problem in this work, specifically focusing on all radiological findings and devices.

3. Data

We sample 60 patients from MIMIC-III for creating this annotated tracking dataset with a total of 638 reports. The average number of reports per patient is 10.6, with the maximum being 33. The reports consist of various imaging modalities including X-ray, computed tomography (CT), CT angiography, magnetic resonance imaging (MRI), and ultrasound as well as different body organs such as chest, head, neck, foot, hip, liver, and kidney. The five most frequent modality types are chest X-ray, CT head, CT abdomen, CT C-spine, and abdomen X-ray. The average length of a report in the collection is 244.7 tokens, with the highest being 1490 tokens. Our dataset includes sufficient radiological linguistic variation as the reports belong to different imaging modalities and describe the imaging interpretation of various anatomies. For annotation, 60 patients are split among three annotators with medical background where each report is annotated by two annotators. We are currently in the process of reconciling the annotations.

4. Annotation Process

We annotate the finding and medical device instances that refer to the same finding/device across reports for a specific patient. Finding here refers to a radiographic finding described in a report. This includes clinical findings (e.g., pneumonia) and imaging observations (e.g., enhancements such as lesion and foci). Device refers to any medical device including tubes and catheters (e.g., endotracheal tube, central venous catheter). We use the Brat tool (Stenetorp et al., 2012) for annotation. The reports of a patient are sorted chronologically using the CHARTTIME attribute of the MIMIC table. Since this is a patient-level annotation, we examine all the sequentially arranged reports for a patient to identify the finding/device instances that correspond to the same finding/device. We assign the same mention identifier to all the entities/mentions across reports that represent the course of a specific finding or device.

4.1. Identifying references of the same finding

The course of a finding can be roughly represented as – (1) initial detection/diagnosis, (2) improved, worsened, etc., and (3) no longer detected. We came up with the following general rules to track a particular finding:

- Identify the first time a finding is detected
- Identify all the other references of the same finding in the subsequent reports highlighting any change in the characteristics of a finding (e.g, a finding may become large, may improve when compared to a previous study etc.)
- Identify all the references until the last report for a patient is reached or if the finding has been resolved

In certain cases, the corresponding location information of a finding serves as a clue in identifying the same reference of a finding across reports. Let us consider the following two examples for a patient:

- Report 1: Questionable ***aneurysm*** at right posterior communicating artery.
- Report 4: Small ***aneurysm*** of size 2.5 mm arises at the origin of posterior communicating artery.

We see that both aneurysms in the two reports are referring to the same aneurysm and hence will be assigned the same mention identifier (belong to the same mention chain). Note that the location *posterior communicating artery* provides a clue that the aneurysms in these reports are discussing about the same finding.

The findings are tracked at the level of the exact anatomical location. This is described through the following points:

1. If the same finding is detected at a different body location or has moved to a different location, we assign different mention identifiers to these findings. For example, ***opacity*** in ***right lower lobe*** is a different finding than an ***opacity*** in the ***left lower lobe***. So different mention IDs will be assigned to these two opacities and are hence part of two different mention chains.
2. We also differentiate findings based on the hierarchical structure of the anatomies. Thus, a ***left frontotemporal fracture*** and a ***skull fracture*** are placed in two different mention chains as the frontotemporal region is a sub-part of the skull.
3. We separate findings based on their laterality information. For example, ***left pleural effusion*** and ***bilateral effusions*** are placed in different mention chains as bilateral indicates that the effusion is also present on the right side.

4. Common finding terms such as ***normal*** and ***unremarkable*** are tracked separately based on the anatomical location or the observation described. For example, the same finding *normal* is placed in separate mention chains corresponding to the two descriptions–'*The appendix is normal.*' and '*Heart size normal.*', as the former is describing about appendix and the latter is about heart size.
5. Again, if the same finding has re-appeared after a period, a different mention identifier is assigned (e.g., *tumor* re-appearing after a few years).

4.2. Identifying references of the same device

Similarly, for tracking the medical devices across reports, the same mention identifier is assigned to entities of a device that represent a specific device. The course can be represented as – (1) insertion, (2) device position status – normal/abnormal, and (3) removal. The following are general rules to track a particular device:

- Identify the first time a device is inserted or placed
- Identify all the other references of the same device in the subsequent reports. This will mainly include updates related to the previously inserted device (e.g., any change in its location or update in the status of its position such as stable, good, satisfactory, unchanged, etc.)
- Continue identifying all the references until the last report for a patient is reached or if the device has been removed

If the same device is re-inserted after a period, we assign a different mention identifier to that device since this indicates a new use of a device. Consider the following sentences from four different reports:

- Report 1: Right ***IJ central venous catheter*** in place and the tip is in distal SVC.
- Report 2: Right ***internal jugular central venous catheter*** in stable position and emanating in middle SVC.
- Report 3: Right ***internal jugular central venous line*** remains in position.
- Report 5: Right ***line*** is terminating in SVC.

Note that all the device mentions in these reports (indicated in bold) are the different variations that are used to refer to the same device and all these mentions are annotated as part of the same mention chain.

4.3. Challenges

The challenges involved in creating this dataset broadly fall under two categories – dependence on context both within and across reports and extensive reliance on radiology domain knowledge. Oftentimes, understanding the context is crucial in correctly annotating the same references of a finding. Table 1 illustrates a scenario where contextual information documented in a long report helps in identifying the coreferences of a finding – subarachnoid hemorrhage. For the first occurrence of hemorrhage, note that linking the *right frontal* location mentioned a few sentences above to the expression–*occupying the immediately subjacent sulci* in the same sentence where hemorrhage occurs indicates that the hemorrhage is associated with right side of the brain. Again, the second occurrence of hemorrhage is associated with the left side as indicated by the location *left frontotemporal* in the same sentence. And the third occurrence is associated with both sides (left and right). Thus, these three instances of hemorrhage belong to three different mention coreference chains.

CT HEAD W/O CONTRAST
Findings:
......
The right frontal fracture is associated with a focal lentiform extra-axial hematoma measuring roughly 8 mm in thickness and 3.5 cm in maximal transverse dimension.
This demonstrates a relatively low-attenuation portion, anteriorly, which may represent acute, non-clotted blood.
This collection may be bounded by the coronal suture, and therefore lie in the epidural space.
There is moderate subarachnoid hemorrhage **occupying the immediately subjacent sulci**, which are slightly flattened, due to the mass effect of the hematoma.
......
No significant extra-axial hematoma is identified at the corresponding left frontotemporal fracture site, though there is subarachnoid hemorrhage **in the sulci in this region**.
......
Impression:
......
Associated subarachnoid hemorrhage at **sites described above**, with possible small associated **right frontal** and **left frontotemporal** contusions
......

Table 1: An example radiology report snippet illustrating the dependence of context for tracking subarachnoid hemorrhage. Findings are in orange, anatomical locations are in green, and the descriptions serving as cues to identify the same finding are **bolded**.

There is also a tremendous dependence on domain knowledge. Table 2 shows a few example sentence pairs (same/across reports) where domain knowledge of different levels are required for annotation. The first example is simple, where *dissociation* and *disruption* are synonymous terms and can be easily identified as coreferences. The second pair is relatively difficult, requiring basic clinical knowledge, with *swelling* and *edema* referring to the same finding entity. The third pair is at a moderate difficulty level, where *atelectasis* and *collapse* belong to the same mention chain and *pneumonia* and *consolidation* belong to another

Difficulty	Example pairs
Simple (synonymous)	Some degree of dissociation as well as lateral displacement of the ossicular chain; Complex fracture of the left temporal bone with evidence of lateral displacement and disruption of the left ossicular chain
Simple (clinical knowledge)	There is a left parietovertex soft tissue swelling; There is extensive left supra- and periorbital soft tissue edema
Moderate (clinical knowledge)	There is new patchy opacity at the left lung base, which may represent resolving postoperative atelectasis with effusion, but *pneumonia* cannot be excluded; New retrocardiac collapse/*consolidation* and bilateral effusions
Complex (clinical knowledge)	There is mild prominence of the pulmonary vascular markings without overt evidence for failure; In the interval, there is increased interstitial edema and small-moderate bilateral pleural effusions.

Table 2: Examples denoting reliance on domain knowledge for annotation.

Item	Count
Avg no. of reports per patient	10.6
Total reports	638
Avg no. of tokens per report	244.7
Min no. of mention chain per patient	8
Max no. of mention chain per patient	110
Total mention chains	2292
Total singleton mention chains	1102
Longest chain length	53
Avg chain length (excluding singletons)	4
Avg no. of tokens per mention	1.44
Total entities (radiological finding)	4978
Total entities (medical device)	894

Table 3: Dataset statistics.

mention chain. The fourth example requires a deeper knowledge where *vascular markings* and *interstitial edema* refer to the same finding.

4.4. Annotation Statistics

Some basic statistics of our annotated dataset are shown in Table 3. We highlight the five most frequent finding and device mentions in Table 4 (note that *"tip"* is a mention that is often documented while referring different medical devices). In terms of inter-annotator agreement, the overall F1 agreement for annotating the mention spans (considering exact span match) is 0.55. The disagreements are mainly related to selecting certain modifier terms describing a radiological findings (e.g., selection of the span *"free intraabdominal air"* by one annotator and only *"air"* by another). For coreference resolution, we calculate the inter-annotator agreement using MUC and CoNLL F1 metrics, and the values are 45.24 and 42.1, respectively.

Finding	Count	Device	Count
effusion	398	tip	144
pneumothorax	238	ng tube	103
fracture	229	endotracheal tube	101
opacity	180	chest tube	42
atelectasis	176	swan-ganz catheter	36

Table 4: Top five frequent mentions in the dataset.

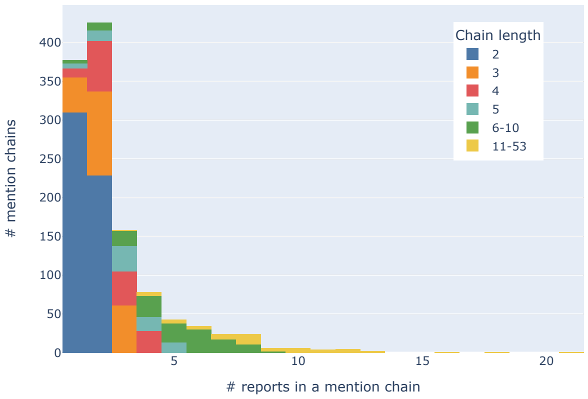


Figure 2: Coverage of reports in mention chains. The x-axis indicates the number of different reports of a patient covered in a mention chain whereas the y-axis indicates the actual number of mention chains.

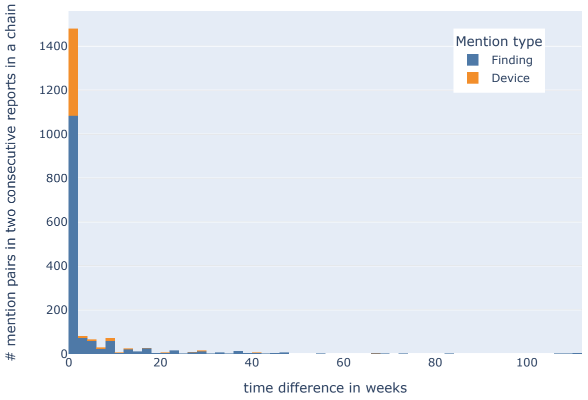


Figure 3: Time difference between two mentions annotated in two consecutive reports in a chain. Each bin denotes an interval of two weeks.

We provide more insights about our annotated corpus through Figures 2, 3, and 4. Figure 2 illustrates the number of different reports that are included while annotating the mention chains. Each stack in a bar highlights the proportion of mention chains according to their lengths (i.e., # mentions in a chain). It is interesting to note that there are more mention chains of length 2 where only a single report contains both the men-

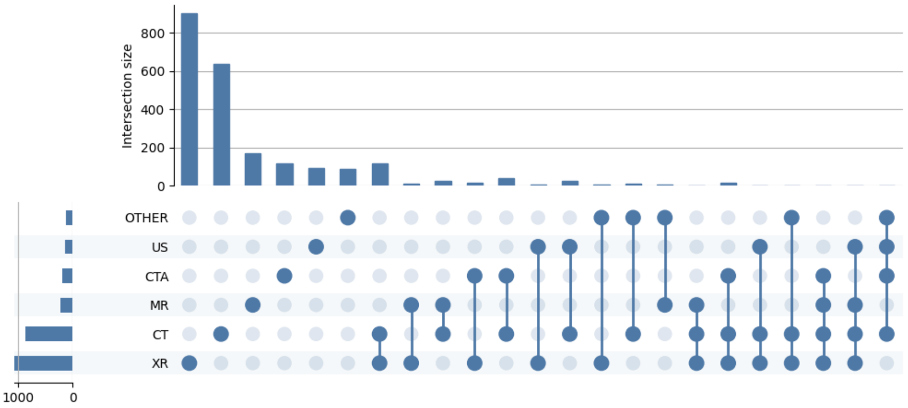


Figure 4: Distribution of imaging modalities in mention chains. XR - X-ray, CT - Computed Tomography, MR - Magnetic Resonance, CTA - CT Angiography, US - Ultrasound, OTHER - other modalities.

tions than when the two mentions are present in two different reports (represented in blue). Also, the number of chains of lengths 3 and 4 are the highest when the chains contain mentions from only two reports.

We show the distribution of temporal distances (in weeks) between co-referring mentions in two sequentially ordered reports of a patient in Figure 3. Overall, more radiological findings are co-referred than medical devices in radiology reports. We observe that the majority of the coreferences between two consecutive reports occurred within an interval of 2 weeks whereas the maximum interval was found to be 2.15 years.

We also illustrate the overlap of imaging modalities of the reports in the annotated mention chains using the UpSet visualization technique (Lex et al., 2014) in Figure 4. While a majority of the mention chains contain mentions described in either only X-ray or CT reports, we do find the inclusion of mentions described in multiple modalities. Among two-modality combinations, (X-ray, CT), (CT, CTA), (CT, MR), and (CT, Ultrasound) are the most frequent co-occurring modalities. Among three-modality combinations, (X-ray, CT, CTA) is the most frequent. We also see a very small percent of mention chains spanning four modalities.

5. Methods

We frame the tracking task as a cross document coreference resolution (CDCR) problem. We apply two baseline methods for automatically identifying the coreferences of findings and devices across all radiology reports of a patient. First is a simple string matching-based baseline, whereas in the second we employ a BERT-based classification approach to predict the mention chains. Since CDCR is the focus of this work, we use the gold mentions.

5.1. Rule-based baseline

We perform sentence segmentation and word tokenization using NLTK. We combine all entities or mentions at a patient level. Then all possible mention pairs are generated. If the lower-cased version of the two mention strings in a pair match, we consider that these two mentions will belong to the same chain. All these mention pairs are then combined to construct the chain.

5.2. BERT-based baseline

In this approach, given a mention pair, we use BERT as a binary classifier to predict whether the two mentions are coreferences. Specifically, we apply BERT in a sentence pair classification setting where information about the two mentions are combined to form the input sequence. Later, the output generated by BERT for all mention pairs corresponding to a patient is combined to predict the final mention chains. We describe the details in the following sub-sections.

5.2.1. Pre-processing

First, we generate all possible mention pair combinations for each patient. We then generate positive and negative pair instances for fine-tuning BERT using gold mention chain information. Since there is imbalance in the number of positive and negative instances (negative instances being 25 times as many positive instances), we randomly sample negative instances such that there are equal instances of positive and negative pairs.

While forming the input sequence to BERT, we provide additional contextual information associated with the two mentions besides the mention spans. We incorporate anatomy and radiology modifier information surrounding a mention span in the sequence. This is grounded on the point that two finding mentions with the same name (e.g., fracture) are placed in separate chains based on their different anatomical locations (e.g., skull vs hip) or different associated modifiers (e.g., right vs left). For this, we leverage the Stanza python library (Qi et al., 2020) and use the clinical model package for identifying the observation, anatomy, and their corresponding identifiers. Specifically, we apply the radiology named entity recognition (NER) model (Zhang et al., 2021) that was trained on radiology reports from three hospitals using a bidirectional LSTM character-level language model. We feed in the pretokenized text generated from NLTK to the Stanza NER pipeline.

5.2.2. Fine-tuning BERT

We fine-tune a BERT$_{\text{LARGE}}$ model to classify whether the two mentions in a pair are co-referring. We initialize the model parameters obtained by pre-training BERT on MIMIC-III clinical notes (Si et al., 2019). We frame our mention pair classification problem as a text pair classification task. First, we use only the mention spans of the two mentions to construct the BERT input sequence as: [CLS]$m1$[SEP]$m2$[SEP], where $m1$ and $m2$ are the spans of the two mentions in a pair. Next, to provide additional information to the BERT model about both the mentions in a pair, we encode the anatomies as well as the anatomy and observation modifiers predicted by Stanza in the sentences containing the mentions. Following the standard BERT input format used in text pair classification configuration, we separate the information corresponding to the two mentions using the special [SEP] token, where anatomy and modifier information of each mention are delimited by a comma. Specifically, we include the Stanza-generated anatomy and modifiers in the left and right of a mention with an window size 5 in the order they appear in a sentence. We construct the BERT input sequence as follows for a mention pair:

[CLS]$m1$, $anty_i(m1)$, ...$anty_n(m1)$, $mod_i(m1)$, ...$mod_n(m1)$[SEP]$m2$, $anty_i(m2)$, ...$anty_n(m2)$, $mod_i(m2)$, ...$mod_n(m2)$[SEP]

Here, $anty_i(m1)$ refers to all the anatomy terms surrounding mention $m1$. Similarly, $mod_i(m2)$ refers to all the modifier terms surrounding mention $m2$.

The output corresponding to the [CLS] token is used to classify if the two mentions are co-referring. The BERT classifier output is then processed to generate the mention chains. All the pairs for which BERT predicted as coreference positive are merged to form the coreference chains. Further, the predicted chain information is converted to CoNLL format for evaluation.

6. Evaluation

We evaluate the methods using gold mentions. We perform 5-fold cross validation to evaluate the performance of the BERT-based approach for CDCR. For each of the 5 iterations, our dataset of 60 patients are split into training, validation, and test sets in the ratio of 60, 20, and 20 %, respectively. The BERT classifier is applied to all possible mention pairs in the test sets. We report the results using the CR evaluation metrics– MUC, B^3, CEAF$_e$, the average F1 of these metrics i.e., CoNLL F1, and BLANC. MUC (Vilain et al., 1995) is a link-based evaluation metric that is based on the minimum number of coreference links required to translate from gold to predicted mention chains. B^3 (Bagga and Baldwin, 1998) is a mention-based metric where the evaluation uses the recall or precision of the individual mentions. For each mention in the gold chains, B^3 recall considers the fraction of the correct mentions that are included in the predicted chain containing that mention. The main assumption of CEAF (Luo, 2005) is that each gold chain should be mapped to only one response chain, and vice versa. BLANC (Recasens and Hovy, 2011; Luo et al., 2014) is another link-based metric where the recall and precision are calculated by averaging the recall and precision of coreference and non-coreference links.

Model	P (%)	R (%)	F1	Acc
Mentions	44.83	85.89	58.91	95.53
+ Context	52.76	86.3	65.49	96.61

Table 5: 5-fold CV results of BERT$_{\text{LARGE}}$ models for classifying if two mentions in a pair are coreferring. P - Precision, R - Recall, Acc - Accuracy.

We use the BERT$_{\text{LARGE}}$ cased model to classify the mention pairs. The model is pre-trained on MIMIC-III notes for 320K steps. We set the maximum sequence length at 128, learning rate at $2e$-5, and the number of training epochs at 4.

7. Results

We show the results of our BERT classification models in Table 5. We illustrate a few sample errors of the BERT classifiers in Table 6. In most of the false positive cases, we observe that the mention strings are the same and better learning of more broad context is required. The false negative errors indicate the need to incorporate more domain-specific knowledge. We then use the output of the BERT models to perform coreference resolution across reports. The cross-report coreference resolution results of the string matching baseline as well as both the BERT variants are in Table 7. We use the gold mention spans in this evaluation. Although the BERT classifier that uses context performs better than the one that uses only mention spans (as per the performance measures in Table 5), we see that the CDCR performance of the latter is better for all metrics. We also observe that the recall values of MUC, B^3, and BLANC are higher for the BERT (mentions) model than the string-matching method that has better precision values (the case is reverse for CEAF$_e$).

8. Discussion

We create an annotated cross-document coreference resolution (CDCR) dataset in the radiology domain to track the same radiological findings and medical devices across all reports of a patient and apply BERT-based baseline method to perform CDCR. The task of CDCR is relatively under-explored in the clinical domain, and in this work we propose a sufficiently large dataset with an average of 10.6 reports per patient (compared to previous 3 notes per patient in (Wright-Bettner et al., 2019)). Additionally, this is the first CDCR dataset in radiology.

Mention pairs	Corresponding sentences	Category	Reason
NG tube; NG tube	**Report-5** Compared with prior radiograph, an NG tube has been withdrawn and there is significant dilatation of the colon lying just below the right hemidiaphragm; **Report-10** An *NG tube* terminates with its tip in the stomach	FP	More deep understanding of context is required (e.g., *"withdrawn"* in Report-5 indicates that the NG tube in Report-10 is different from the first one). Sufficient contextual information is not incorporated into the models
thrombus; thrombus	**Report-3** The grayscale ultrasound of the veins of the upper extremities demonstrated filling defect in the right cephalic vein at the level of the antecubital fossa consistent with *thrombus*; **Report-13** No intraluminal *thrombus* is identified		
collapse; atelectases	**Report-1** There is increased retrocardiac density, consistent with left lower lobe *collapse* and/or consolidation; **Report-20** There is cardiomegaly with *atelectases* in the left upper lobe as well as atelectasis in the left lower lobe.	FN	More domain knowledge understanding is required to link the correlated findings
hemorrhage; hematoma	**Report-1** There is no intraparenchymal *hemorrhage* identified; **Report-6** There is a small left frontal subdural *hematoma*, slightly larger than prior CT studies		

Table 6: Common error types of BERT classification models. FP - False Positive, FN - False Negative.

Methods	MUC			B^3			$CEAF_e$			CoNLL	BLANC		
	P	R	F1	P	R	F1	P	R	F1	F1	P	R	F1
String match	80.18	70.36	74.95	83.52	70.75	76.6	61.88	76.46	68.4	73.32	78.52	71.6	74.44
BERT (mentions)	68.56	95.31	79.65	40.84	86.57	55.23	76.53	23.46	35.84	56.91	53.39	77.18	50.29
BERT (mentions + context)	67.46	92.58	77.87	32.72	85.7	46.48	76.12	18.97	29.99	51.45	49.79	67.92	39.13

Table 7: CDCR performances. Precision - P %, Recall - R %. 5-fold cross validation results are reported for BERT models.

The results in Tables 5 and 7 indicate that there is enough scope for performance improvement. A brief analysis of the output from the BERT classifiers suggests that incorporating rich radiology-specific domain knowledge will be useful in improving CDCR systems. For example, there is potential in encoding knowledge about relations between different human anatomies, knowledge about clinical correlation between various radiological findings (e.g., 'consolidation' and 'pneumonia'), and information about findings that are more often coreferred across different imaging modalities. Another promising avenue is allowing the model to learn more broad cross-report context (e.g., by leveraging certain language patterns in the reports suggesting any potential coreference such as 'compared to previous study'). We also intend to investigate the impact of BERT classifier output on the various CDCR evaluation metrics in detail.

An interesting method to explore for CDCR model development using this annotated dataset is by adopting the recently proposed cross-document language modeling technique that uses a new pre-training approach that has shown to be effective for several multi-document downstream tasks including CDCR and multihop question answering (Cattan et al., 2021a). The pre-training technique considers two main ideas: pre-training over sets of multiple related documents and usage of dynamic global attention pattern over masked tokens. We intend to use this pre-training approach and develop a CDCR system similar to the CDCR pairwise scoring framework proposed in a recent work (Caciularu et al., 2021). Here, we can feed the whole radiology reports corresponding to the two mentions in a pair into the CDLM rather than feeding only the local context of the mentions (e.g., surrounding words of a mention). We also aim to build an end-to-end CDCR system where the predicted mention spans are used to infer the mention chains instead of the gold mentions, although this relies on a robust extraction system to identify the radiological entities accurately (which is oftentimes challenged by the presence of different modifier terms described in conjunction with the main finding terms). From the clinical application perspective, we also aim to extend this dataset to cancer domain that demands long-term tracking of findings such as tumor and cyst in the future.

9. Conclusion

We construct a new cross-document coreference resolution (CDCR) dataset for tracking radiological findings and devices across reports. Our annotated dataset comprises of 638 radiology reports belonging to 60 patients. This resulted in a total of 2292 mention chains. We provide a detailed description of our annotation process and demonstrate some important aspects of the dataset including the major challenges (both from the annotation and model development perspectives). We apply two baseline methods to automatically identify the cross-report coreferences. The system performances are low to moderate, and we plan to leverage this annotated dataset to develop more advanced methods for radiology CDCR in our later work.

Acknowledgments This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB: R21EB029575), the Patient-Centered Outcomes Research Institute (PCORI: ME-2018C1-10963), and the Cancer Prevention and Research Institute of Texas (CPRIT RP210045).

10. Bibliographical References

Apostolova, E., Tomuro, N., Mongkolwat, P., and Demner-Fushman, D. (2012). Domain Adaptation of Coreference Resolution for Radiology Reports. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 118–121.

Bagga, A. and Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Barhom, S., Shwartz, V., Eirew, A., Bugert, M., Reimers, N., and Dagan, I. (2019). Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189.

Bozkurt, S., Alkim, E., Banerjee, I., and Rubin, D. L. (2019). Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm. *Journal of Digital Imaging*, 32(4):544–553.

Bradshaw, T., Weisman, A., Perlman, S., and Cho, S. (2020). Automatic image classification using labels from radiology text reports: Predicting Deauville scores. *Journal of Nuclear Medicine*, 61(supplement 1):1410–1410.

Bugert, M., Reimers, N., and Gurevych, I. (2021). Generalizing Cross-Document Event Coreference Resolution Across Multiple Corpora. *arXiv:2011.12249 [cs]*.

Caciularu, A., Cohan, A., Beltagy, I., Peters, M., Cattan, A., and Dagan, I. (2021). CDLM: Cross-Document Language Modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662.

Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., and Dagan, I. (2021a). Cross-document Coreference Resolution over Predicted Mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107.

Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., and Dagan, I. (2021b). Realistic Evaluation Principles for Cross-document Coreference Resolution. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 143–151.

Cattan, A., Johnson, S., Weld, D., Dagan, I., Beltagy, I., Downey, D., and Hope, T. (2021c). SciCo: Hierarchical Cross-Document Coreference for Scientific Concepts. *arXiv:2104.08809 [cs]*.

Datta, S. and Roberts, K. (2022). Fine-grained spatial information extraction in radiology as two-turn question answering. *International Journal of Medical Informatics*, 158:104628.

Datta, S., Ulinski, M., Godfrey-Stovall, J., Khanpara, S., Riascos-Castaneda, R. F., and Roberts, K. (2020). Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports. *LREC ... International Conference on Language Resources & Evaluation : [proceedings]. International Conference on Language Resources and Evaluation*, 2020:2251–2260.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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*, pages 4171–4186.

Hassanpour, S. and Langlotz, C. P. (2016). Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*, 66:29–39.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992.

Luo, X., Pradhan, S., Recasens, M., and Hovy, E. (2014). An Extension of BLANC to System Mentions. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2014:24–29.

Luo, X. (2005). On Coreference Resolution Performance Metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Mabotuwana, T., Hall, C. S., Tieder, J., and Gunn, M. L. (2018). Improving Quality of Follow-Up Imaging Recommendations in Radiology. *AMIA Annual Symposium Proceedings*, 2017:1196–1204.

Mabotuwana, T., Hall, C. S., Hombal, V., Pai, P., Raghavan, U. N., Regis, S., McKee, B., Dalal, S., Wald, C., and Gunn, M. L. (2019). Automated Tracking of Follow-Up Imaging Recommendations. *AJR. American journal of roentgenology*, pages 1–8.

Miller, T., Dligach, D., Bethard, S., Lin, C., and Savova, G. (2017). Towards generalizable entity-centric clinical coreference resolution. *Journal of Biomedical Informatics*, 69:251–258.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Recasens, M. and Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Rubin, D. L., Willrett, D., O'Connor, M. J., Hage, C., Kurtz, C., and Moreira, D. A. (2014). Automated Tracking of Quantitative Assessments of Tumor Burden in Clinical Trials. *Translational Oncology*, 7(1):23–35.

Si, Y., Wang, J., Xu, H., and Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Son, R. Y., Taira, R. K., and Kangarloo, H. (2004). Inter-document coreference resolution of abnormal findings in radiology documents. *Studies in Health Technology and Informatics*, 107(Pt 2):1388–1392.

Steinkamp, J. M., Chambers, C., Lalevic, D., Zafar, H. M., and Cook, T. S. (2019). Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. *Journal of Digital Imaging*, 32(4):554–564.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Sugimoto, K., Takeda, T., Oh, J.-H., Wada, S., Konishi, S., Yamahata, A., Manabe, S., Tomiyama, N., Matsunaga, T., Nakanishi, K., and Matsumura, Y. (2021). Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729.

Syeda-Mahmood, T., D, P., Wong, K. C. L., D, P., Wu, J. T., D., M., H, M. P., Jadhav, A., D, P., Boyko, O., and D, M. D. P. (2020). Extracting and Learning Fine-Grained Labels from Chest Radiographs. *arXiv:2011.09517 [cs]*.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *MUC*.

Wood, D., Guilhem, E., Montvila, A., Varsavsky, T., Kiik, M., Siddiqui, J., Kafiabadi, S., Gadapa, N., Busaidi, A. A., Townend, M., Patel, K., Barker, G., Ourselin, S., Lynch, J., Cole, J., and Booth, T. (2020). Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). In *Medical Imaging with Deep Learning*.

Wright-Bettner, K., Palmer, M., Savova, G., de Groen, P., and Miller, T. (2019). Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 1–10.

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., and Langlotz, C. P. (2021). Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.