Natural Language Processing for Lines and Devices in Portable Chest X-Rays

Journal:	*AMIA 2010 Annual Symposium*
Manuscript ID:	AMIA-1508-A2009.R1
Manuscript Type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Rubin, Daniel; Stanford University, Radiology Wang, Dan; Department of Veteran's Affairs; Medcisive LLC Chambers, Dallas; VAPAHCS, GRECC Chambers, Justin; VA Palo Alto Health Care System South, Brett; VA Palo Alto Health Care System Goldstein, Mary; VA Palo Alto Health Care System, Geriatrics Research Education and Clinical Center (GRECC)
Primary Axis Classification:	II.13. Measuring/improving patient safety and reducing medical errors

Natural Language Processing for Lines and Devices in Portable Chest X-Rays

Daniel Rubin, MD, MS,[1,2] Dan Wang, PhD,[1] Dallas A. Chambers,[1] Justin G. Chambers,[1] Brett R. South, MS,[3] Mary K. Goldstein, MD, MS[1, 2]
[1]VA Palo Alto Health Care System, Palo Alto, CA; [2]Stanford University, Palo Alto, CA; [3]VA Salt Lake City Health Care System, IDEAS Center, Salt Lake City, UT

Abstract

Radiology reports are unstructured free text documents that describe abnormalities in patients that are visible via imaging modalities such as X-ray. The number of imaging examinations performed in clinical care is enormous, and mining large repositories of radiology reports connected with clinical data such as patient outcomes could enable epidemiological studies, such as correlating the frequency of infections to the presence or length of time medical devices are present in patients. We developed a natural language processing (NLP) system to recognize device mentions in radiology reports and information about their state (insertion or removal) to enable epidemiological research. We tested our system using a reference standard of reports that were annotated to indicate this information. Our system performed with high accuracy (recall and precision of 97% and 99% for device mentions and 91-96% for device insertion status). Our methods are generalizable to other types of radiology reports as well as to other information extraction tasks and could provide the foundation for tools that enable epidemiological research exploration based on mining radiology reports.

Introduction

Portable chest X-ray (CXR) imaging is a prevalent examination for evaluating critically ill patients. Patients in the intensive care unit (ICU) often have a variety of medical devices inserted as part of their course of clinical care. These patients frequently have complications such as infections that could be related to the presence and length of time that medical devices are present. Complications are known to be caused by medical devices such as blood-borne infections that increase the cost of medical care and cause morbidity.[1, 2] Such complications could be studied by analyzing the devices present and the length of time they are present. Such information about these devices could be extracted from the medical record through natural language processing (NLP) of radiology reports to determine the "dwell time" of devices, based on detecting mentions of devices in the reports and their "insertion status"—when they were inserted or removed. Detecting insertion status is possible using NLP because radiologists generally describe the appearance or disappearance of devices seen on CXR in the radiology reports.

Discovering how device types or device dwell time correlate with important clinical parameters requires two tasks in NLP: (1) identifying mentions of devices in CXR reports, and (2) extracting information related to the mention of the device indicating whether it has been inserted, removed, or remains present compared with prior studies.

A variety of NLP systems have been developed in radiology and biomedicine, including systems to identify named entities such as drugs, genes, and diseases; NLP systems have been used to identify key phrases or named entities followed by the application of linguistic rules for associating the discovered entities to extract semantic information from free-form text.[3, 4] However, a system that also determines the temporal context—when the device was inserted or removed—has not yet been developed.

Our goal in this work is to develop an NLP system to identify medical devices in CXR reports and to infer the status of the insertion of the device in the report (whether it has been inserted, removed, or remains in place). We report on our work to create this system, the Chest X-Ray Device Extractor (CXDE). In this paper we describe our preliminary work in creating CXDE, a preliminary evaluation of the methodology, and directions for future work, including using this system to enable epidemiological research in large databases of text reports.

Methods

Data

We randomly selected 566 reports of portable CXR exams obtained in the ICU at 46 Veterans Affairs Medical Centers from across the country during 1/1/2006 to 1/1/2009. Randomization was done by: (1) stratifying by medical center and (2) stratifying by length of stay (LOS) in the ICU. LOS groups included short LOS (1-3 days), medium LOS (4-7 days), and long LOS (8-10 days). The sample was taken at a 40:40:20 ratio of short, medium, and long LOS, respectively, equally distributed across all medical centers.

NLP System

Our CXDE system analyzes each report in two sequential steps, implemented as processing resources in the GATE framework.[5]

1. Segmentation and annotation: The first processing step segments the text into individual sentences, and for each sentence, identifies particular annotations of interest related to medical devices. We used the ANNIE pipeline in GATE with customizations to the ANNIE Gazetteer to analyze each sentence and to create annotations indicating the presence of devices. The ANNIE Gazetteer provides a dictionary lookup facility that matches words or phrases to determine the appropriate "lookup type" of annotation. We added 5 custom lookup types, one indicating *devices* that are mentioned in the report, and 4 types of terms indicating the insertion status: *insertion* of a device, continued *presence* of a device, *removal* of a device, and *absence* of a device. We will refer to word or phrase that matches a lookup type the "term" and the lookup type as its "type." For example, "IJ catheter" is a term and *device* is its type; "in satisfactory position" is a term (an insertion status) and *presence* is its type.

For the list of device terms, we included single words like "tube," phrases such as "internal jugular catheter," and abbreviations like "CVC" (for Central Venous Catheter), and "IABP" (for Intra-Aortic Balloon Pump). The complete list of device terms were obtained from three sources: UMLS, Radlex,[6] and portable CXR reports from 40 VA sites spread over the U.S. geographically.

To deduce the list of terms indicating insertion or removal of devices, we relied on expert clinician input and from searching actual portable chest X-Ray reports. Terms or phrases indicating recent insertion include "inserted", "interval placement", and "has been placed." Terms or phrases indicating recent removal include "removed," "no longer seen," or "interval withdrawal." The largest collection was the terms or phrases indicating the presence of a device; such as "redemonstrated," "remains", "in satisfactory position," and "no change in position."

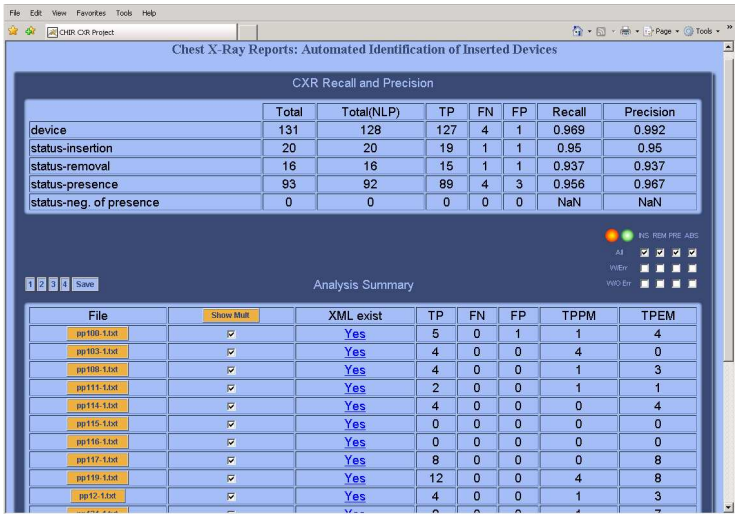
2. Inference to deduce insertion status. The second processing step analyzes the annotations derived from first step to infer the insertion status of devices identified in the report. The algorithm assumes that if a device mention occurs within the same sentence as a term annotated as an *insertion-status* (and meets proximity-based rules described below), then the *insertion-status* applies to that device. The insertion status of that device is then directly inferred from the particular type of *insertion-status* assigned to the device (*recent insertion*, *presence*, *recent removal*, and *absence*). Thus, in the sentence, "Since previous exam, the internal jugular catheter remained in satisfactory position," the CXDE inference step would identify the *device* term "internal jugular catheter" and the *insertion status* term, "in satisfactory position" (annotating its insertion status as *presence*), and it would then infer that the latter insertion status applies to the former device (i.e., that *device* = internal jugular catheter" and *insertion status* = "present").

The proximity between device mentions and insertion status is quantified as the number of tokens between them. A set of proximity rules is evaluated to infer whether the insertion status applies to the device. The basic rule is that if the *insertion status* term is separated from a *device* by 6 or fewer tokens, the association between the *device* and the *insertion status* terms is made. Alternate inferences based on proximity are evaluated in three specialized situations.

1) Clause Boundaries Trumps Proximity: if there is a clause boundary (the presence of a comma) between the device and the insertion status, then no association is made between them. For example, in the sentence: "Nasogastric tube redemonstrated, the IJ catheter in the main pulmonary artery removed" there are two insertion-status terms that could be associated to IJ catheter, "redemonstrated" and "removed". The former is closer to the device term than the latter term, but the CXDE will exclude "redemonstrated" as the insertion-status term for IJ catheter because there is a clause boundary between the device and the insertion status terms. As a result, the insertion-status term "removed" would be correctly identified.

2) Precedence Order of Statuses: if more than one insertion status term is mentioned within the proximity window of a device, CXDE chooses the term to use based on a precedence order according to the type of insertion status. The precedence ordering defines that terms of type *removal* or *insertion* take precedence over terms of type *presence*. For example, the sentence "Interval placement of the left subclavian line is seen with its tip projecting over the cavoatrial junction" contains insertion-status terms "Interval placement" and "seen." Both are within the maximum proximity limit; however, the CXDE would assign the term "Interval placement" to "subclavian line" by precedence ordering, correctly inferring that the subclavian line was recently inserted.

3) Assumption of nearest device status: if device mentions are conjoined using "and" then each will be considered for association with the insertion status in that sentence. Consider the sentence: "The endotracheal tube and nasogastric tube are in satisfactory position." Based solely on proximity, the device term "nasogastric tube" would be associated with "in satisfactory position." Since "endotracheal tube" is conjoined with another *device* in the same sentence, it receives this same insertion status ("*presence*"). This rule is also applied when there are lists of devices associated with the same status.



	Total	Total(NLP)	TP	FN	FP	Recall	Precision
device	131	128	127	4	1	0.969	0.992
status-insertion	20	20	19	1	1	0.95	0.95
status-removal	16	16	15	1	1	0.937	0.937
status-presence	93	92	89	4	3	0.956	0.967
status-neg. of presence	0	0	0	0	0	NaN	NaN

File	Show Mult	XML exist	TP	FN	FP	TPPM	TPEM
pp100-1.txt	☑	Yes	5	0	1	1	4
pp103-1.txt	☑	Yes	4	0	0	4	0
pp108-1.txt	☑	Yes	4	0	0	1	3
pp111-1.txt	☑	Yes	2	0	0	1	1
pp114-1.txt	☑	Yes	4	0	0	0	4
pp115-1.txt	☑	Yes	0	0	0	0	0
pp116-1.txt	☑	Yes	0	0	0	0	0
pp117-1.txt	☑	Yes	8	0	0	0	8
pp119-1.txt	☑	Yes	12	0	0	4	8
pp12-1.txt	☑	Yes	4	0	0	1	3

Figure 1. CXDE Debugger showing scoring summary

Annotation Schema and Guidelines

To develop a reference standard for testing the performance of CXDE, three members of the research team independently annotated 90 reports randomly selected from the 566-report corpus. We developed an annotation schema to record the location of devices mentioned in texts as well as the type of insertion status described for each of them. We used the Knowtator text annotation tool[7] to develop the annotation schema and to provide a tool the readers used to record the text annotations. The annotation schema comprised four classes: Device/Line, Device/Line Status, Laterality, and Device/Line Quantity. Information from only the former two classes was used to evaluate CXDE.

The reports were divided into 6 batches of 15 reports each and annotations were done by two readers, with the third serving as the adjudicator for the batch. The role of adjudication rotated among the staff for each batch. All questions that arose during the adjudication process were resolved by the study investigators.

An annotation guideline document was developed by the study investigators, which provided detailed instructions to the annotators as to the task, requirements, and examples. All readers studied this document and resolved any questions prior to undertaking the annotation task.

Refinement of CXDE

The CDXE system was developed by studying and performing testing in approximately 300 reports (randomly selected from the 566 report corpus, and excluding any reports in the reference standard). We developed the rules and Gazetteer term lists by iterative refinement, comparing CXDE results with annotations in the reference standard. To enable reasonable efficiency in training the CXDE, we developed a web-based tool, the CXDE Debugger (Figure 1). The CXDE Debugger enables semi-automatic comparison of Knowtator output results with CXDE results and detailed analysis of the comparison for a large number of reports. For each report, the CXDE Debugger computes the number of True Positive (TP) matches, False Negatives (FN), and False Positives (FP) for the 5 categories: devices, insertion, removal, presence, and absence of the device. The summary page shows the aggregated results from all reports in that run (Figure 1). It is also possible to drill down into each report (Figure 2) to view the complete report text. Individual annotations from Knowtator or from CXDE are highlighted with separate colors. Using the CXDE Debugger,

Figure 2. CXDE Debugger: single report view showing highlighting of identified terms. Green = CXDE only; Orange= Knowtator only; Yellow = Exact Match of CXDE and Knowtator.

Table 1. Precision and Recall in 90 reports using CXDE

Class	Total(Standard)	Total(NLP)	TP	FN	FP	Recall	Precision
Device Mention	148	144	143	5	1	0.966	0.993
Status-Insertion	23	22	21	2	1	0.913	0.954
Status-Removal	16	16	15	1	1	0.937	0.937
Status-Presence	107	106	102	5	4	0.953	0.962

deficiencies in the CXDE or errors in manual annotation or problems with ambiguous text can be rapidly identified, which is essential for improving the accuracy of the CXDE.

Evaluation of CXDE

The annotations in the reference standard were used as the gold standard against which the CXDE was evaluated. CXDE processed the 90 reports in the reference standard and recorded all mentions of devices and classified each according to category of insertion status. These data were compared with the annotations in the reference standard and precision and recall metrics were calculated as:

$$\text{recall} = TP / (TP + FN)$$

$$\text{precision} = TP / (TP + FP)$$

where TP=count of true positive, FN = count of false negatives, and FP=count of false positives. The recall measures how well the CXDE is able to identify all instances (device mentions and insertion status) in the reference standard, while precision measures how likely an instance found by CXDE is an actual instance. If there is a partial text match between the CXDE and the gold standard (an overlap in text spans), it was counted as a TP. For example, if the CXDE found a device term, "IJ line" but the gold standard contained the text "double lumen IJ line" this was counted as a TP. For insertion statuses, a TP occurred as long as the inference was correct for the device term referenced; there does not need to be overlap between the text spans of the insertion-status term. For example, in the sentence "There is an endotracheal tube 2 cm above the carina," the insertion-status term indicating presence could be "There is" or "2 cm above the carina" or even the device term itself. While the inference of being present is clear, there can be multiple, equally valid, insertion-status terms. This is particularly true for the status-presence type.

Results

Table 1 shows the absolute counts and precision/recall for each of 5 types of annotation (device mentions and 4 types of insertion status) detected by CXDE and the reference standard. There were 148 device terms identified. Of these devices, 23 were identified as being inserted, 16 being removed, 107 being present, and non absent in the report text. The number of true positive (TP), false negative (FN), and false positive (FP) classifications are also shown in addition to the precision and recall for each classification. There were no examples of device insertion status of absence (either in the reference standard or detected by CXDE) so this type of insertion status is omitted from Table 1.

As can be seen in the Table, the CXDE identifies device mentions with a recall and precision of 97% and 99%, respectively, and identifies the *presence* status type with a recall and precision of 95% and 96%, respectively. For the *inserted* and *removed* status type, the recall and precision are about 91-94% and approximately 94%, respectively. These are reasonably good performance values.

Discussion

A major source of data relevant to epidemiological research is the radiology report, which provides information about patient conditions and complications. This information is recorded in unstructured text, and thus NLP methods are needed to access and analyze large numbers of reports. The VA recently established the Consortium for Healthcare Informatics Research (CHIR), the goal of which is to change free text in the electronic medical record into structured data. The current work has been undertaken as part of the CHIR program with the vision of enabling researchers to analyze radiology reports to answer questions such as finding out whether particular devices or their dwell time correlates with infection or treatment response.

Our CXDE system shows promising results for the task of detecting device mentions in radiology reports and for determining the context of those mentions (i.e., whether the device has been inserted, removed, or remains in place). There was excellent agreement between CXDE and the reference standard in our study, likely due to the relative ease of identifying device mentions in CXR reports. On the other hand, determining the insertion status of the device is more challenging, and our approach to inferring that information was accurate as well (Table 1).

The CXDE system was developed using 300 CXR reports, and it was evaluated using 90 reports (none

of which were used for developing the system). While it would have been better to use larger numbers of reports, our results are sufficiently strong to suggest that these results would be similar with larger corpora. Few if any NLP systems such as CXDE can become reasonably reliable or accurate without significant refinement to enhance the basic rules and key terms. CXDE has attained the present level of recall and precision ranging from 91% to 99% after training with about 300 reports from over 46 VA sites distributed nationally across the US.

Trick and colleagues previously described a system to detect mentions of central venous catheters.[8] Our work differs from theirs in that our system detects a wide range of medical devices in radiology reports as well as determining their insertion state.

There are limitations of our study. Our approach to inferring device insertion status does not take into account negations or hedges, which could confound the validity of the inference. We will be adding processing modules to our NLP module to recognize negation and hedge qualifications that modify the text describing the insertion status. A second limitation is that our method assumes that all device mentions and insertion status occur within the same sentence. Co-references among devices and their insertion status could certainly occur across sentences and we currently do not account for that. Refinements to our method to include co-references will be undertaken in the future. A third limitation is that our method is heuristic, and not based on a formal language analysis. While the latter could be more robust, it would have been more costly and time-consuming to develop. Our approach is simple and can be extended by adding rules to the CXDE processing pipeline. Finally, while we acknowledge that CXDE has deficiencies, we believe they can be overcome by identifying specific sentence patterns where the system is deficient and making system modifications to address those issues.

A benefit of the CXDE system is that it is automated and built using a sample of CXR reports from many different institutions. Thus, we believe it will likely continue to perform well on future CXR reports. We plan to continue development and to test future versions with a larger number of CXR report.

Conclusion

We developed an NLP system for detecting device mentions in radiology reports and for determining their insertion status. The current performance of the system appears sufficient to be used in use cases such as epidemiological research where detecting devices and determining their dwell time is important. As such NLP methods become more commonplace and refined, we expect clinicians to be routinely mining unstructured patient data to improve their practice and the quality of care.

Acknowledgements

We thank Shuying Shen for designing the scheme for chart selection. Views expressed are those of the authors and not necessarily of the Department of Veterans Affairs. This study was carried out after receiving appropriate IRB approvals from relevant VA entities. This study was undertaken as part of the VA HSR&D CHIR Translational Use Case Projects HIR 09-007 (PI Goldstein), a component of CHIR HIR 08-374 (PI Samore).

References

1. Pittet D, Tarara D, Wenzel RP. Nosocomial bloodstream infection in critically ill patients. Excess length of stay, extra costs, and attributable mortality. JAMA. 1994 May 25;271(20):1598-601.
2. Renaud B, Brun-Buisson C. Outcomes of primary and catheter-related bacteremia. A cohort and case-control study in critically ill patients. Am J Respir Crit Care Med. 2001 Jun;163(7):1584-90.
3. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc. 2006:269-73.
4. Goldin IM, Chapman WW. Learning to Detect Negation with 'Not' in Medical Texts. ACM SIGIR '03 Workshop on Text Analysis and Search for Bioinformatics, Association for Computing Machinery; 2003; Toronto, Canada. 2003.
5. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA; 2002.
6. Rubin DL. Creating and curating a terminology for radiology: ontology modeling and analysis. J Digit Imaging. 2008 Dec;21(4):355-62.
7. Ogren PV. Knowtator: a protégé plug-in for annotated corpus construction. Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Morristown, NJ, USA: Association for Computational Linguistics; 2006. p. 273–5 %W New York, New York.
8. Trick WE, Chapman WW, Wisniewski MF, Peterson BJ, Solomon SL, Weinstein RA. Electronic interpretation of chest radiograph reports to detect central venous catheters. Infect Control Hosp Epidemiol. 2003 Dec;24(12):950-4.