Toward Complete Structured Information Extraction from Radiology Reports using Machine Learning

Jackson M. Steinkamp, Hospital of the University of Pennsylvania; Darco Lalevic; Charles Chambers, MCIT; Hanna Zafar, MD, MHS; Tessa Cook, MD, PhD, CIIP

Introduction
Unstructured and semi-structured radiology reports represent an underutilized trove of information for machine learning (ML)-based clinical informatics applications, including abnormality tracking systems, research cohort identification, point-of-care summarization, semi-automated report writing, and as a source of weak data labels for training image processing systems. Clinical ML systems must be **interpretable** to ensure user trust. To create interpretable models applicable to all of these tasks, we can build general-purpose systems which extract all relevant human-level assertions or “facts” documented in reports; identifying these facts is an information extraction (IE) task. Previous IE work in radiology has focused on a limited set of information, and extracts isolated **entities** (i.e., single words such as “lesion” or “cyst”) rather than complete “facts”, which require the linking of multiple entities and modifiers. Here, we develop a prototype system to extract **all** useful information in a radiology report (findings, recommendations, clinical history, procedures, imaging indications and limitations, etc.), in the form of complete, contextualized facts. We construct an information schema to capture the bulk of information in reports and develop real-time ML models to extract this information.

Hypothesis
It is feasible to develop an ML system for complete information extraction from radiology reports.

Methods
We developed a schema of fact types based on the content of our institution’s abdominopelvic radiology reports (see Table 1 for examples). Each fact type has exactly one “anchor” entity (e.g., finding, diagnosis, procedure, recommendation) and any number of modifiers and relations (e.g., size of finding, statement of change over time, indication for procedure, diagnostic interpretation, etc.) represented by nearby text spans, which we call information “slots”. The anchor entity and its associated slots constitute a complete fact. We manually labeled 100 abdominopelvic radiology reports with their complete factual content. We developed a neural network model to convert a report into a series of discrete facts, using custom-trained FastText vectors for word representations and a two-step detection process: 1) detect anchor entities, 2) identify surrounding text spans corresponding to associated information slots. We evaluated our model’s performance on an unseen test set of 20 reports.

Results
Our information schema consisted of 17 fact types with 70 information slot types. The schema successfully represented the informational content in our radiology reports: 98% of the raw text was included within at least one fact. In total, 4,478 facts were labeled with 14,082 information slots. Across all fact types, micro-averaged F1 score was 80.7% (recall 77.4%, precision 84.2%) for detecting anchor entities and 74.1% (recall 71.6%, precision 76.8%) for detecting associated slots.

Conclusion
We demonstrate the feasibility of complete real-time information extraction from radiology reports, using a small corpus of abdominal reports and no external knowledge bases. More training data is likely to further improve the system.
Statement of Impact
Downstream applications include summarization, tracking of abnormal findings, intelligent chart search, research cohort identification, semi-automated expansion of ontologies, and creation of weak labels for large imaging data sets.

Keywords
machine learning, natural language processing, information extraction, radiology reports

Table 1:
Examples of some fact types in the information schema, with true labels. Within each row, colors in the parsed example match the colors in the "anchor" entity and associated slots columns. Slots in black are unused in the example.

| Fact Type                          | "Anchor" entity | Associated slots                                                                 | Example instance of fact                                                                 |
|------------------------------------|-----------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| Radiologist asserts imaging finding | Finding         | Location, observation modifier, timing, image citation, size measurement, description of change over time, diagnostic interpretation | Stable septated inferior interpolar cystic lesion is seen on image 1/4, measuring 7 x 7 mm. |
| Patient has diagnosis              | Diagnosis       | Timing of diagnosis, severity, previous workup, previous therapeutic measure      | 62-year-old male with history of SMA dissection status post repair.                     |
| Patient has/had procedure          | Procedure       | Time of procedure, body location, indication, outcome/consequence, where/by whom procedure performed | There are post-surgical changes related to prior ileal resection.                       |
| Patient has/had imaging study      | Imaging modality| Time of occurrence, anatomic region, indication, sequence obtained, details about contrast, summary of findings | CT of the chest performed with this exam will be separately evaluated.                  |
| Patient has follow-up recommendation | Recommendation | Desired timing, indication                                                        | Recommendation of MRI on nonemergent basis to evaluate the renal lesion.               |
| This study has a limitation         | Limitation      | Reason for limitation                                                             | Evaluation of the bowel is hampered by the lack of oral contrast.                      |
| Slots usable with all fact types    |                 | Negation of fact, uncertainty about fact                                          |                                                                                       |