Pruning the Hedges: Using Natural Language Processing to Extract Meaningful Information from Radiology Reports

Daniel J. Goff, MD, PhD, University of California, Davis; Thomas Loehfelm, MD, PhD

Hypothesis

A natural language processing (NLP) pipeline accurately identify and extract relevant diagnoses from standard radiology reports.

Introduction

Medicine is a field where information overload has become reality. Within the electronic medical record a single patient's chart may contain thousands of notes, images, diagrams, and other data. Sifting through this information is daunting, yet necessary if one is to clearly understand a patient's clinical situation. Radiologists too are faced with this problem, relying on the chart and prior imaging reports to provide context with which to interpret imaging studies.

Unfortunately, in the busy world of modern practice there is often little time to delve into the chart., and our lack of knowledge of the patient's history surely interferes with our ability to make important diagnoses. In the midst of such massive information, tools must be developed to help us parse through complex information systems and efficiently extract need-to-know information.

Natural language processing (NLP) is a field focused on interaction and translation of human language and computational language. Its uses are broad, including speech, relationship extraction, and automated summarization. Importantly, NLP algorithms can analyze large data sets and can quickly extract meaningful language. These chunks of language, known as concept unique identifiers (CUIs), are linked to formal frameworks for organizing language. In Medicine these might include diagnoses, treatments, or history. In theory, a well-developed NLP pipeline could scan through a patient’s chart, identify salient clinical information, and create a coherent summary.

With this ultimate goal in mind we developed an NLP pipeline to extract diagnostic information from radiology reports. Using a database of standard clinical terminology (the systemized nomenclature of medicine), and an automated NLP annotator (cTakes), we analyzed a subset of radiology reports and measured how accurately our automated pipeline could extract diagnoses.

Methods

Two radiologists used the Brat Rapid Annotation Tool (version 1.3) (2) to manually annotate asserted and negated pathology concepts in the Impression sections of 50 radiology reports (Figure 1). Reports were randomly selected from all contrast-enhanced CT Abdomen/Pelvis exams performed at our center between July and December 2016. Discrepancies in the annotation set were resolved by mutual agreement to reach a consensus, gold standard set of annotations.
All reports were processed by the Clinical Text Analysis Knowledge Extraction System (cTakes), an open-source natural language processing tool for clinical documents (1), using a slightly modified version of a default pipeline (AnnotatePlaintextFastUMLSPreprocessor). The reports were first manually reviewed to extract headers, which cTakes used to identify the relevant subsections of the report. All DiseaseDisorderMentions and SignSymptomMentions identified in the Impression section of the report were stored to a local MySQL database. In the case of overlapping annotations of the same class, the result with the maximum length was retained and the other(s) discarded. For example, if presented with the phrase "Acute appendicitis", cTakes will annotate "appendicitis" and "acute appendicitis" each as DiseaseDisorderMentions - we would only store "acute appendicitis".

Interobserver agreement was considered perfect if the radiologists chose the same start, end, and negation status of the concept, or partial if the radiologists used overlapping but non-identical annotation frames with the same negation status. Disagreement was assigned when only one radiologist annotated the concept, or when different negation status was assigned to otherwise perfect or partial agreements. All cases of disagreement were resolved by consensus to produce the final gold standard set of manual annotations.

The gold standard annotation set was then compared to the cTakes annotation set. True positives were counted when cTakes identified perfect or partial matches compared to the gold standard. False positives were counted when cTakes annotated concept that was not included in the gold standard. False negatives were counted when cTakes annotated a concept from the gold standard but with incorrect polarity, or when cTakes failed to annotate a concept from the gold standard set. Our methods did not allow for calculation of true negatives. Sensitivity, precision, and accuracy were thus determined assuming zero true negatives.

Results

Interobserver Agreement
For all 50 report Impressions reviewer 1 annotated 164 concepts and reviewer 2 annotated 149 concepts. In total, accounting for concepts annotated by one reviewer and not the other, there were 186 annotated concepts. There was perfect agreement on 104/186 annotations (56%) and partial agreement on an additional 24 annotations, leading to an overall agreement rate of 69% before consensus discussions.

26 annotations (14%) were discarded during the consensus process as not relevant, leaving 160 annotations in the gold standard set. 15 annotations were discarded because they were made to sections other than the Impression, and the remainder because they were anatomy- or procedure-related terms, not disease processes. Consensus was reached after one round of discussion.
cTakes Performance
cTakes found perfect matches for 93 of the annotations in the gold standard set (93/160; 58%), and partial matches for an additional 40 (25%), yielding a true positive rate of 133/160 (83%) (Figure 2). cTakes assigned the incorrect polarity to 7 otherwise perfect matches and 3 otherwise partial matches (10/160; 6%). For 10 of the 40 partial matches, cTakes actually did a better job than the human annotators by identifying a more specific concept (e.g. cTakes annotated "multiple pulmonary nodules" or "partial small bowel obstruction" while the humans simply annotated "pulmonary nodules" or "small bowel obstruction"). 8 partial matches were due to cTakes splitting a contiguous phrase into two separate annotations (most often AnatomicSiteMention + DiseaseDisorderMention, e.g. adrenal + nodule, pulmonary + nodule, omental + mass). cTakes found 40 additional DiseaseDisorderMentions and SignSymptomMentions in the impression sections which were not included in gold standard annotation set (false positives).

Summary
There were 133 true positive matches, 27 false negative matches, and 40 false positive matches, yielding a sensitivity of 83%, precision of 77%, and accuracy of 67%. If we assume that each word in the impression that is not annotated by cTakes or by the human reviewers is a true negative annotation, then there are 1918 true negatives, which allows specificity to be determined (98%) and increases accuracy to 97%. The F1 score, the harmonic mean of precision and sensitivity, is 80%.

Discussion
The overall performance of cTakes for annotating relevant pathology terms from our radiology reports was good, and allows for a reasonable summary of the impression to be automatically generated by considering the polarity of identified concepts. For example, the following report impression:

Impression:
1. Right nephrostomy tube is in place in appropriate position. No hydronephrosis.
2. Right greater than left renal cortical scarring.
3. Postoperative changes status post ventral abdominal hernia repair, with mesh, with stable 6 cm postoperative seroma in the left superficial abdominal wall.
4. Hepatosplenomegaly mildly dilated main portal vein, suggestive of portal hypertension.

Can be summarized from cTakes annotations as:
1. Nephrostomy
2. No hydronephrosis
3. Scarring
4. Abdominal hernia
5. Repair
6. Postoperative seroma
7. Hepatosplenomegaly
8. No portal hypertension

Note that the polarity of the portal hypertension annotation was assigned incorrectly.

An automatically generated summary composed of concepts from structured ontologies can be useful in clinical practice. Upon starting a new case, for example, a radiologist could be presented with a list of anatomically relevant concepts identified in all of the prior studies for the patient without having to launch and read each report separately. Real-time named entity recognition incorporated into dictation software could recognize terms as they are described and review historical studies on the same patient to assist in determining interval changes, or review studies from different patients to identify other patients that had similar findings. This process of automatic patient cohort identification is a powerful application of clinical NLP and is currently used to facilitate large scale clinical trials (3, 4), but to our knowledge has not been incorporated at the point of care to facilitate the radiologist's workflow.

Our preliminary study has several limitations. We developed a small set of gold standard annotations, and only for CT Abdomen/Pelvis exams. Also, we only considered pathology concepts. Adding procedural and anatomic concepts, as well as grammar and syntax, would allow for a deeper level of natural language understanding, at the cost of increased complexity. Our preliminary data suggest that allowing for compound annotations (AnatomicalSiteMention + DiseaseDisorderMention) would increase the specificity and utility of the annotations. At this point, the negation detection algorithms are probably not accurate enough to produce reliable auto-generated summaries, but the summaries may still be a useful adjunct to manual report review.

Conclusion

Using a simple NLP pipeline we successfully extracted relevant diagnoses from a cohort of radiology studies with 67% accuracy, comparable to our interobserver agreement of 69%. We plan to further refine these methods and expand them to cover all diagnostic imaging studies to allow accurate and automatic report summaries available at the point-of-care workstation.

References


Keywords

NLP, Natural Language Processing, named entity recognition, cTakes, annotation