Title

Automated Radiology Report Summarization Using an Open-Source Natural Language Processing Pipeline

Journal

Journal of Digital Imaging, 31(2)

ISSN

0897-1889

Authors

Goff, Daniel J
Loehfelm, Thomas W

Publication Date

2018-04-01

DOI

10.1007/s10278-017-0030-2

Peer reviewed

J Digit Imaging (2018) 31:185–192
https://doi.org/10.1007/s10278-017-0030-2

Automated Radiology Report Summarization Using an Open-Source Natural Language Processing Pipeline

Daniel J. Goff[1] · **Thomas W. Loehfelm**[1]

Published online: 30 October 2017

Abstract Diagnostic radiologists are expected to review and assimilate findings from prior studies when constructing their overall assessment of the current study. Radiology information systems facilitate this process by presenting the radiologist with a subset of prior studies that are more likely to be relevant to the current study, usually by comparing anatomic coverage of both the current and prior studies. It is incumbent on the radiologist to review the full text report and/or images from those prior studies, a process that is time-consuming and confers substantial risk of overlooking a relevant prior study or finding. This risk is compounded when patients have dozens or even hundreds of prior imaging studies. Our goal is to assess the feasibility of natural language processing techniques to automatically extract asserted and negated disease entities from free-text radiology reports as a step towards automated report summarization. We compared automatically extracted disease mentions to a gold-standard set of manual annotations for 50 radiology reports from CT abdomen and pelvis examinations. The automated report summarization pipeline found perfect or overlapping partial matches for 86% of the manually annotated disease mentions (sensitivity 0.86, precision 0.66, accuracy 0.59, F1 score 0.74). The performance of the automated pipeline was good, and the overall accuracy was similar to the interobserver agreement between the two manual annotators.

Keywords NLP · Report summarization · Data extraction · Radiology report

✉ Thomas W. Loehfelm
twloehfelm@ucdavis.edu

[1] Department of Radiology, University of California Davis Health System, 4860 Y Street, Suite 3100, Sacramento, CA 95817, USA

Background

Radiology reports are a potential treasure trove of data that document the presence or absence, time course, and treatment-response of diseases. They describe in detail the imaging features of specific disease entities and at their best consolidate the varied findings into a concise and coherent impression statement that forms the basis for downstream clinical decision-making.

The information in radiology reports exists as free-text and therefore in its raw form is mostly inaccessible to data-mining approaches. A variety of natural language processing (NLP) techniques have been applied to extract data from and classify radiology reports and other clinical texts (reviewed in [1]). Most NLP techniques involve constructing a pipeline or a series of processing steps that are sequentially applied to input text to yield some output. In developing a new NLP technique, it is necessary to manually annotate a set of reports so that the output of the NLP technique can be compared to this "gold-standard" set of manual annotations. Once the NLP technique has been developed and validated in this way, it can then be deployed to process more reports without any further manual annotation required. In the case of a report classifier, those processing steps attempt to divide a collection of reports into a limited set of discrete classes—CT chest reports may be divided into those that assert the presence of acute pulmonary embolism and those that do not, for example, and in this way, the classifier can be used to identify a cohort of patients that share specific clinical characteristics. Report classifiers are generally developed for a specific task, and while they usually perform well at that task, they must be retrained for each separate task (classifying CT head reports on the presence or absence of intracranial hemorrhage or ultrasound abdomen reports on the presence of absence of acute cholecystitis, etc.).

In contrast to a report classifier, which takes as its input a free-text report and produces as its output a discrete class assignment for the report as a whole, a data-extraction pipeline takes free-text input and produces structured data as output. The specific data elements and their structure vary depending on the project and are defined upfront by the NLP pipeline developer. These may be project-specific, such as BI-RADS descriptors and categories from mammography reports, or they may be more general, such as *all* measurements, anatomic site mentions, and disease entities from any report.

Most prior work using NLP on radiology reports has focused on extracting a narrow set of features from a limited set of report types. Bozkurt et al. reported high accuracy in extracting lesions and their BI-RADS characteristics from mammography reports [2]. Pham et al. developed a machine learning algorithm that had excellent accuracy to detect thromboembolic disease (DVT and PE) from angiography and venography reports but was considerably less accurate at the more general task of detecting clinically relevant incidental findings [3]. More recently, Hassanpour and Langlotz reported impressive results extracting general features from a diverse report set using a machine learning approach to annotate reports based on an information model capturing anatomy, observations, modifiers, and uncertainty, and determined that this approach is superior to "non-machine learning" approaches [4].

NLP is an inherently complex task, due to ambiguities and individual stylistic differences in free-text reports that are relatively simple for a human reader to understand in context but that are challenging to reduce to logic statements that can be understood computationally. Two approaches to NLP in the clinical domain can be generalized to those that are rule-based and those that are statistical or rely on machine learning.

The rule- and dictionary-based approaches borrow from general NLP systems (i.e., those not confined to the biomedical domain) and work by first identifying general grammar concepts such as parts of speech, noun phrases, and coreference and pronoun resolution. Specific categories of words (nouns, for example) are then reduced to their base forms and then mapped to a predefined dictionary of terms. The advantage of a rule-based system is that it follows predictable logic and is therefore comprehensible and tunable. In addition, since the output is directly mapped to existing dictionaries of terms (examples include SNOMED-CT, CPT, or ICD-10), the additional layers of information that are encapsulated in these structured ontologies are accessible to the NLP pipeline and output. Key disadvantages of rule-based approaches are that they rely on the reports following predictable rules of grammar, which is not always the case with clinical texts, and on the completeness of the reference dictionaries.

Statistical NLP approaches follow the same initial steps of identifying general grammar concepts but then use statistical inferences gleaned from a gold-standard manually annotated corpus to extract meaningful relationships in the input text. To illustrate the difference between rule-based and statistical approaches to NLP, consider an NLP project to catalog the location of abscesses identified on CT. Consider a sentence such as "There is an abscess in the space of Retzius." In a rule-based approach, if "space of Retzius" is not in the reference dictionary, it will not be flagged as an anatomic space and this occurrence might be overlooked by a rule-based pipeline. A statistical NLP approach, on the other hand, might recognize that the phrase "There is an abscess in" is always followed by an anatomic location, and so even without knowing beforehand that "space of Retzius" is an anatomic location, a statistical NLP tool can assume that it is and flag it appropriately.

Most modern clinical NLP systems use a hybrid approach, utilizing statistical or machine learning approaches for pipeline components that benefit from that flexibility and utilizing rule-based approaches including dictionary lookup where the additional layers of information available in the reference ontology is useful. The open-source Apache project Clinical Text Analysis and Knowledge Extraction System (cTAKES) is one such hybrid tool. cTAKES uses machine learning for some components, such as the dependency parser, coreference resolver, and relation extractor [5, 6], and rule-based dictionary lookup for named-entity recognition [7].

An accurate hybrid NLP system could quickly extract the most salient information from a patient's previous radiology reports, a task that currently falls upon the radiologist. Not only is this task time-consuming, but in the modern high-volume radiology department, there is a very real risk of overlooking critical studies and/or data. With these issues in mind, our goal is to develop a robust tool to accurately summarize radiology reports by identifying disease entities mentioned in the impression section of radiology reports. Towards that end, we developed a custom NLP pipeline assembled from existing cTAKES components and measured its performance against manual parsing and annotation, the current gold standard.

Methods

Study Setting

The Institutional Review Board approved this HIPAA-compliant study. The requirement for informed consent was waived due to the retrospective nature of the work. Two radiologists (one body fellowship trained attending with 2 years of experience and one PGY4 resident) reviewed 50 reports randomly selected from all contrast-enhanced CT abdomen/pelvis exams performed at a single academic radiology site between July and December 2016 (Table 1).

Table 1 Summary and demographic information of the 50 randomly selected CT abdomen and pelvis reports

Number of reports	50
Number of radiologists represented	14 faculty; 22 residents
Number of reports written by the study authors	11 (TWL 9, DJG 2)
Number of patients represented	50
Patient age	2–81 years old; average 51.2 years
Patient sex	25 males; 25 females
Patient disposition	22 emergency, 21 outpatient, 7 inpatient

Gold-Standard Annotations

The two reviewers used an open-source text annotation tool (BRAT rapid annotation tool, version 1.3 [8]) to manually annotate asserted and negated disease entities in the impression sections of the reports (Fig. 1). Discrepancies were resolved by consensus to yield a gold-standard set of annotations.

NLP Pipeline Construction

cTAKES version 4.0.1 was installed in the Eclipse Integrated Development Environment. An aggregate processing pipeline was assembled from stock cTAKES components using the UIMA Component Descriptor Editor. The pipeline consists of a report segmenter, sentence detector, tokenizer, lexical variant generator, context-dependent tokenizer, part-of-speech tagger, dependency parser, chunker, lookup window annotator, UMLS dictionary lookup annotator, and a series of assertion annotators that identify history of assertions, subject, polarity (i.e., negation status), uncertainty, and disease severity modifiers. Many of the components are general text processing tools and are relatively self-explanatory (e.g., segmenter, sentence detector, tokenizer, part-of-speech tagger, and chunker). The lexical variant generator is a component that generates canonical forms of words (e.g., reduces plural nouns to the singular form) and is typically included in pipelines that include dictionary lookup. The context-dependent tokenizer considers the immediate context of some tokens and thereby can differentiate a numeric token that is a component of a *date* from a numeric token that is a component of a *measurement*. A lookup window annotator sets bounds on portions of a sentence that are then used by downstream components. For example, dictionary lookup might only be performed on tokens that have been tagged as *nouns* by the part-of-speech tagger, so the lookup window might be set to *noun phrases* only in order to minimize processing of parts of the sentence that are unlikely to contain useful information. By default, cTAKES uses the entire sentence as the lookup window. The dictionary lookup annotator is the main component of the pipeline that annotates the report text with concepts identified in the source dictionary. The report segmenter used regular expressions based on extracted report section headers (e.g., *Impression*), which were manually defined based on local report template design prior to processing. The lookup window annotator was configured to save only the longest text span in case of overlapping annotations of the same class. For example, in the phrase "acute cholecystitis," cTAKES will identify "acute cholecystitis" and "cholecystitis" as separate disease mentions—we stored the longest text span ("acute cholecystitis") and discarded the other ("cholecystitis").

Text Processing

The 50 radiology reports were processed with the automated-processing pipeline (Fig. 2). The output annotations were stored to a local database developed specifically for this project. For each identified annotation, the report accession number, report section, start and end positions of the identified text span in the source text, annotation type (based on the cTAKES type system), polarity, UMLS concept unique identifier (CUI), history of determination, subject (patient vs. family member), and uncertainty indicator were stored to the database.

Interobserver Agreement and Comparison to the Gold Standard

Interobserver agreement was considered perfect if the radiologists chose the same start, end, and polarity of the annotated concept, and partial if the radiologists used overlapping but not identical annotation frames with the same polarity. Disagreement was assigned when only one radiologist annotated the concept or when different polarity was assigned to otherwise perfect or partial agreements.

Annotator performance was assessed by comparing cTAKES output to the consensus set of manual annotations. True positives were counted when cTAKES identified perfect or partial matches compared to the gold standard. False positives were counted when cTAKES annotated a concept that was not included in the gold standard. False negatives were counted when cTAKES annotated a concept from the gold standard but assigned the incorrect polarity or when cTAKES failed to annotate a concept from the gold-standard set. See Table 2 for an example of how annotations were accounted for. Our methods do not allow calculation of true negatives. Sensitivity, precision, and accuracy were thus determined assuming zero true negatives.

Assuming that the manual annotators identified 100% of the relevant findings, our post hoc analysis of our sample size of 160 annotations indicates power = 0.80 and alpha = 0.95 to detect a 5% difference in performance between cTAKES and manual annotators.

Performance was assessed on the raw dataset as well as a version containing only the unique annotations, which were

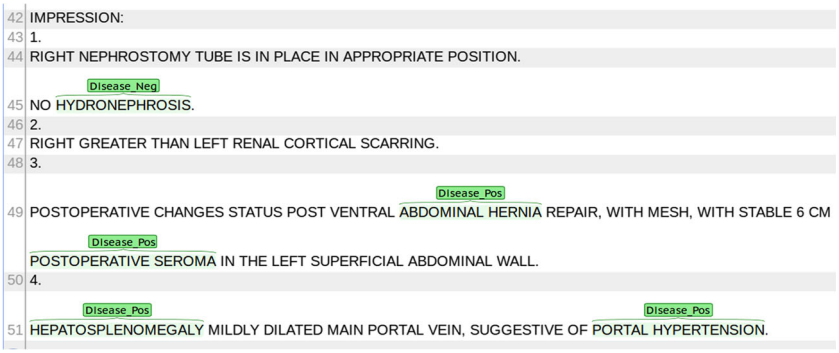

Fig. 1 Example manual annotation using the BRAT rapid annotation tool. For each report, disease mentions were identified in the *Impression* section and annotated as Disease_Pos or Disease_Neg as indicated by the text. Fifty reports were annotated separately by two different radiologists

derived from the raw annotation dataset by applying the *Remove Duplicates* function of Microsoft Excel.

Statistics

All statistical analyses were performed in R (version 3.3, R Foundation for Statistical Computing, Vienna, Austria).

Results

Interobserver Agreement

Reviewer 1 annotated 164 concepts and reviewer 2 annotated 149 concepts from the 50 reports. 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.

Twenty-six annotations (14%) were discarded during the consensus building process as irrelevant, leaving 160 annotations in the manual gold-standard set. Of the discarded annotations, 15 were determined to be in sections other than *Impression*, and the remainder were discarded because they were anatomy- or procedure-related terms, not disease processes. Consensus was reached after a single round of discussion.

cTAKES Performance

In total, cTAKES identified 249 strings, 128 disease/disorder mentions, and 121 sign/symptom mentions, from 10 different UMLS type unique identifiers (TUIs, Table 3). We excluded annotations in the TUI categories T041 (mental process; $n = 2$) and T184 (sign or symptom; $n = 16$) due to their unacceptably high false-positive match rate, leaving a total of 231 strings: 128 disease/disorder mentions and 103 sign/symptom mentions. Of these, 93 were perfect matches to the manual annotations in the gold-standard set (93/160; 58%) and 44 were partial matches (44/160; 28%), yielding a true-positive rate of 137/160 (86%, Table 4).

cTAKES assigned the incorrect polarity to 6 otherwise perfect matches and 4 otherwise partial matches (10/160; 6%). For 8 of the 44 partial matches, cTAKES identified a more specific concept (e.g., cTAKES annotated "multiple pulmonary nodules" and

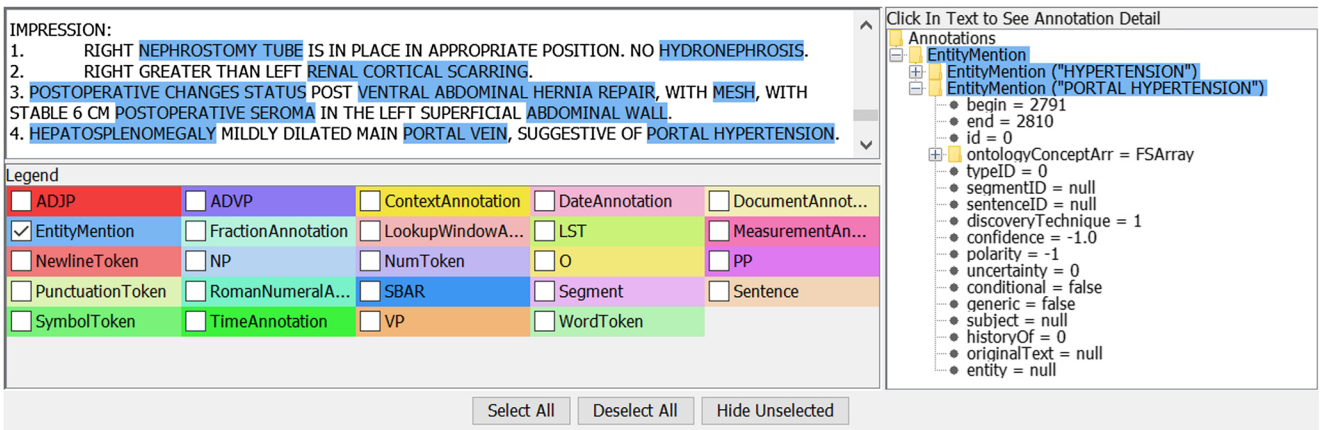

Fig. 2 Automated annotation using cTAKES. Left: reports were analyzed for EntityMentions, which include both disease states and anatomical terminology. Right: each EntityMention was processed using the custom NLP pipeline to identify several different language properties. These language properties include polarity (presence versus absence of the entity), subject (who does the entity pertain to?), history (is this a current or a prior condition?), and uncertainty (are we confident that this is the correct entity?)

Table 2 Example report *Impression* section, with corresponding manual and cTAKES annotations. There are three manual annotations (italicized) and four cTAKES (three in the signs/symptoms type class and one in the disease/disorder type class). We count two of these as true positives: cTAKES found a perfect match for "pulmonary nodules" and a partial match for "improving atelectasis"; one false negative: "fluid collection" from the manual annotations set not flagged by cTAKES; and one false positive: cTAKES annotated "follow-up" as a sign/symptom, but this has no overlap with any manual annotation. Note that cTAKES generated two overlapping annotations, "pulmonary nodules" as a sign/symptom, and "nodules" as a disease/disorder. Accounting for this can get confusing. It is not accurate to include this string as a true positive since the overlapping phrase "pulmonary nodules" is already counted as a true positive. Including "nodules" would artificially inflate the number of true-positive matches. It is also not fair to include it as a false positive, since it is a partial match to a gold-standard annotation. Finally, it is not fair to count it as a false negative in the context of the manual annotation task, which did not ask the manual annotators to differentiate such semantic nuances. We excluded these by defining false-positive matches as only those cTAKES matches that had *no overlap* with a gold-standard annotation, without regard to polarity of the match. A match with incorrect polarity is already accounted for as a false negative

Example *Impression* section:

1. Indeterminate fluid collection in the left lower quadrant.
2. 9 and 5 mm pulmonary nodules at the left lung base, now better visualized due to improving atelectasis. Recommend 3 month follow-up.

Manual annotations:

1. Indeterminate *fluid collection* in the left lower quadrant.
2. 9 and 5 mm *pulmonary nodules* at the left lung base, now better visualized due to *improving atelectasis*. Recommend 3 month follow-up.

cTAKES, automated annotations:
Sign/symptom type class

1. Indeterminate fluid collection in the left lower quadrant.
2. 9 and 5 mm *pulmonary nodules* at the left lung base, now better visualized due to improving *atelectasis*. Recommend 3 month follow-up.

Disease/disorder type class

1. Indeterminate fluid collection in the left lower quadrant.
2. 9 and 5 mm pulmonary *nodules* at the left lung base, now better visualized due to improving atelectasis. Recommend 3 month follow-up.

"partial small bowel obstruction" while the human annotators annotated "pulmonary nodules" and "small bowel obstruction"). Eleven partial matches were due to cTAKES splitting a contiguous phrase into two separate annotation classes (most often AnatomicSiteMention + DiseaseDisorderMention, e.g., adrenal + nodule, pulmonary + nodule, omental + mass). cTAKES found 71 annotations that were not included in the gold-standard annotation set (false positives).

In summary, there were 137 true-positive matches, 23 false negatives, and 71 false positives, yielding a sensitivity of 0.86, precision of 0.66, and accuracy of 0.59. The F1 score, the harmonic mean of precision and sensitivity, is 0.74 (Table 5).

A number of concepts were repeated several times, either in the same report or across multiple different reports. For example, the word "diverticulitis" occurred four times in our corpus, and "mass" occurred six times. In order to distinguish the baseline performance of the annotation system from the confounding effects of term frequency, we also calculated performance metrics for the set of unique strings in the manual annotation set. There were 117 unique terms in the manual set. cTAKES found 108 true-positive matches, 9 false negatives, and 44 false positives, yielding a sensitivity of 0.92, precision of 0.71, accuracy of 0.67, and F1 score of 0.80 (Table 5).

Discussion

The sensitivity of cTAKES for annotating relevant pathology terms from our report corpus was good and allows for a reasonable summary of the *Impression* section of a report to be automatically generated.

Our F1 score of 0.74 compares favorably to the score of 0.85 reported by Hassanpour and Langlotz for their machine learning approach and is significantly better than the F1 score they report for the cTAKES arm of their comparison study (0.58) [4]. The reason for our improved score is likely multifactorial. Hassanpour and Langlotz included some information model classes that were outside the scope of our project (anatomy modifiers and uncertainty assertions). However, it is unlikely that this difference alone explains the different performance. Our concept of "Disease Entities" most closely approximates the "Observation" class in their information model, and on that subset of concepts, they report an F1 score of 0.59, which is still considerably worse than our results. A more likely reason for the difference in performance is that we used the full cTAKES pipeline, including the full SNOMED-CT dictionary, while Hassanpour and Langlotz derived their own dictionary from RadLex and only used the named-entity recognition module from cTAKES. Our results suggest that relying on the full functionality of cTAKES leads to much better performance, and since it is an open-source project, any deficiencies in performance can be addressed through user- and community-level updates. To that end, we used a more recent version of cTAKES than they did, so it is possible that improvements to the software in the time between ours and their study contributed to better performance. The advantage of an open-source system and of using a source dictionary like SNOMED that is incorporated in the UMLS is that concepts identified in SNOMED can easily be cross-referenced to other knowledge sources that add additional layers of information, such as anatomic context for anatomy terms (via the Foundational Model of Anatomy [FMA]), drug class for medications (via RxNorm), and billing and coding information for procedures (via the

Table 3 Summary of cTAKES annotations by type unique identifier (TUI). cTakes identified a total of 249 unique text strings. Of these, 128 were classified as disease/disorder mentions from six different TUIs and 121 were sign/symptom mentions from four different TUIs. Each annotation was compared to the manual annotation list generated by the radiologists (the gold-standard set). A match was assigned if the string was identical to or partially overlapped a term in the manual annotation set. No match was assigned when the cTakes annotation had no overlap with a manual annotation. TUI041 and TUI184 (italics) were excluded from subsequent analyses because of their high false-positive rate

		Disease/disorder mentions		Sign/symptom mentions	
Total number		128		121	
TUI	Definition	Match (TP)	No match (FP)	Match (TP)	No match (FP)
T019	Congenital abnormality	2 (67)	1 (33)		
T020	Acquired abnormality	8 (80)	2 (20)		
T037	Injury or poisoning	11 (73)	4 (27)		
T047	Disease or syndrome	59 (81)	14 (19)		
T190	Anatomical abnormality	12 (86)	2 (14)		
T191	Neoplastic process	11 (85)	2 (15)		
T033	Finding			23 (39)	36 (61)
T041	*Mental process*			*0 (0)*	*2 (100)*
T046	Pathologic function			28 (64)	16 (36)
T184	*Sign or symptom*			*1 (6)*	*15 (94)*
Total		103 (80)	25 (20)	52 (43)	69 (57)
Total (excluding T041 and T184)		103 (80)	25 (20)	51 (50)	52 (50)

Current Procedural Terminology [CPT] code set). In our opinion, this additional information and access to open-source tools to access it outweighs the cost of slightly diminished performance compared to more proprietary tools.

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. It would allow the radiologist to selectively review studies that mention specific disease entities, rather than reviewing each study first to determine *if* it is relevant. Real-time named-entity recognition incorporated into the dictation workflow could recognize terms as they are described and review historical studies on the same patient to assist in assessing interval changes or even review studies from other patients to identify a cohort with similar imaging 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 [9, 10] but, to our knowledge, has not been incorporated at the point-of-care to facilitate the radiologist's workflow and accurate diagnosis.

Table 4 Summary of positive and negative matches between the manual annotation set and the automated annotation set. There were a total of 160 manual annotations (gold-standard set) and 231 cTAKES annotations. True positives were defined as manual annotations with an exact or partial match in the cTAKES annotation set. False negatives were defined as manual annotations with no cTAKES match or those with a match that was assigned incorrect polarity by cTAKES. False positives were defined a cTAKES annotation that had no overlap with a manual annotation

	Number	Percentage
Manual annotations (gold standard)	160	
Automated annotations (cTAKES)	231	
True positives	137	85.6
Perfect cTAKES matches	93	58.1
Partial cTAKES matches	44	27.5
False negatives	23	14.4
Manual annotation with no cTAKES annotation	10	6.3
Incorrect polarity assigned by cTAKES	13	8.1
False positives	71	31
cTAKES annotation with no manual annotation	71	31

Our study has several limitations. We developed a small set of gold-standard annotations, only for a single exam type. We only considered disease entity concepts. Adding procedural and anatomic concepts, as well as more robust 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.

Another limitation is the high false-positive rate of cTAKES as applied in this specific project. This is primarily attributable to entities identified under the signs/symptoms category, which account for 49% of the total entity matches but > 70% of the false positives (Table 3). Signs/symptoms entities include a large number of generic modifier words (e.g., mild, worse, likely) and nonspecific terms (e.g., disease,

Table 5 Performance metrics for cTakes automated annotation. Metrics were calculated taking into account all annotations, only unique annotations, and only unique annotations while also ignoring polarity. 95% confidence interval is provided for sensitivity, precision, and accuracy measurements

	All annotations	Unique annotations	Unique, ignoring polarity
True positive	137	102	108
False positive	71	44	44
False negative	23	15	9
Total	231	161	161
Sensitivity	0.86 (0.79–0.91)	0.87 (0.80–0.93)	0.92 (0.86–0.96)
Precision	0.66 (0.59–0.72)	0.70 (0.62–0.77)	0.71 (0.63–0.78)
Accuracy	0.59 (0.53–0.66)	0.63 (0.55–0.71)	0.67 (0.59–0.74)
F1	0.74	0.78	0.80

findings, mass). While not necessarily unimportant, many of these words were not within the scope of our manual annotation process and therefore are not reflected within the gold-standard annotation set. We were able to exclude some false-positive terms by excluding annotations in the UMLS T041 and T184 classes, but other classes contained a more balanced mix of true-positive and false-negative annotations and could not be so easily filtered out. We note that this is not an intrinsic limitation of cTAKES, but more a reflection of the ambiguous language used by radiologists.

Ambiguities of natural language also complicate the task of negation detection. To accurately determine whether a concept is asserted or negated, it is necessary to consider the contextual relationships and dependencies between words in the same sentence and sometimes even between separate sentences. Improving the NegEx negation detection algorithm that cTAKES uses is a field of active research, and we refer interested readers to a recent publication that improves on NegEx by incorporating dependency parsing [11].

The false-positive annotations, those where cTAKES annotated a phrase that was not in the gold-standard set, can be grouped into a number of categories and shed light on the inherent difficulties of the task of natural language processing. Some of the terms that cTAKES annotated are nonspecific phrases used by radiologists in lieu of a more precise term, such as *abnormalities*, *bulge*, *complication*, *disease*, *lesions*, and *thickening*. Other terms that indicate disease course or degree of confidence (*absence*, *definitely present*, *increased size*, *mild*, *probable*, *severe*, and *unchanged*) were annotated by cTAKES separate from whatever disease process they were modifying. The manual annotators flagged the disease entity itself but were not asked to consider these modifying terms for this specific task. These additional terms convey important information to be sure, but they were not flagged by the manual annotators as indicating the true presence or absence of a disease. We recognize that this inherent subjectivity of the manual annotators is a limitation of our work but argue that it is a limitation that is inherent in all natural language processing, including the comprehension that occurs when a person attempts to make sense of a natural language report without the aid of a computer whatsoever.

We think it is important to consider both the full set of annotations, including duplicated terms, as this gives an estimate of true performance in a clinical setting; however, we also think it is valid to assess performance against the unique set of phrases, as this gives an estimate for the baseline function of the annotation tool. Failure to annotate the term "intraperitoneal free air" is a valid mark against the performance of the annotator, but counting *each time* that "intraperitoneal free air" is missed confounds the performance of the annotator itself with the term prevalence.

It is notable that the pipeline metrics are improved by ignoring the polarity of the annotation (Table 5, *unique, ignore polarity*). This highlights the difference in performance between the named-entity detector, which relies on the completeness of the source dictionary, with the performance of the negation detector, which has the more difficult task of considering contextual information. The current negation detection algorithms are not accurate enough to produce reliable automated summaries for direct clinical use, but even partially inaccurate summaries may be a useful adjunct to the current manual report review process. Improved polarity detection would likely lead to substantial improvements in accuracy while maintaining a high degree of sensitivity. With the modular design of cTAKES, as improved components are developed, they can be easily swapped in to the pipeline. Several other refinements could also be introduced to the pipeline that would likely improve the overall performance, such as improved contextual awareness and dependency parsing, taking advantage of the disease severity modifiers that cTAKES already annotates, and incorporating more temporal awareness to define disease progression over time.

Conclusion

In conclusion, a simple NLP pipeline assembled from the open-source cTAKES project components can accurately extract relevant diagnoses from radiology reports. The annotator precision of 66% was comparable to the interobserver agreement rate of 69%, and the F1 score of 0.74 compares favorably with other NLP techniques. We plan to further refine these

methods and expand them to cover all diagnostic imaging studies to allow accurate and automated report summaries to be made available at the point-of-care diagnostic workstation.

References

1. Cai et al.: NLP technologies in radiology research and clinical applications. Radiographics 36(1):176–191, 2016
2. Bozkurt S, Lipson JA, Senol U, Rubin DL: Automatic abstraction of imaging observations with their characteristics from mammography reports. J Am Med Inform Assoc 22(e1):e81–e92, 2015. https://doi.org/10.1136/amiajnl-2014-003009 **Erratum in: J Am Med Inform Assoc. 2015 Sep;22(5):1112.**
3. Pham AD, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, Morello R, Burgun A: Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 15:266, 2014. https://doi.org/10.1186/1471-2105-15-266
4. Hassanpour S, Langlotz CP: Information extraction from multi-institutional radiology reports. Artif Intell Med 66:29–39, 2016
5. Albright D, Lanfranchi A, Fredriksen A et al.: Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc 20:922–930, 2013
6. Zheng J, Chapman WW, Miller TA, Lin C, Crowley RS, Savova GK: A system for coreference resolution for the clinical narrative. J Am Med Inform Assoc 19:660–667, 2012
7. Savova GK, Masanz JJ, Ogren PV et al.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17:507–513, 2010
8. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. In: 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, 2012:102–107
9. Wu ST, Sohn S, Ravikumar KE et al.: Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Ann Allergy Asthma Immunol 111:364–369, 2013
10. Ni Y, Wright J, Perentesis J et al.: Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak 15: 28, 2015
11. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, Beesley C, Dexter P, Max Schmidt C, Liu H, Palakal M: DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform. 54:213–219, 2015