Extracting Actionable Findings of Appendicitis from Radiology Reports Using Natural Language Processing

Bryan Rink\textsuperscript{1}, Kirk Roberts\textsuperscript{1}, Sanda Harabagiu, PhD\textsuperscript{2}, Richard H. Scheuermann, PhD\textsuperscript{2}, Seth Toomay, MD\textsuperscript{2}, Travis Browning, MD\textsuperscript{2}, Teresa Bosler, PMP\textsuperscript{2}, Ronald Peshock, MD\textsuperscript{2}

\textsuperscript{1}The University of Texas at Dallas, Richardson, TX, USA  
\textsuperscript{2}The University of Texas Southwestern Medical Center, Dallas, TX, USA

Abstract
Radiology reports often contain findings about the condition of a patient which should be acted upon quickly. These actionable findings in a radiology report can be automatically detected to ensure that the referring physician is notified about such findings and to provide feedback to the radiologist that further action has been taken. In this paper we investigate a method for detecting actionable findings of appendicitis in radiology reports. The method identifies both individual assertions regarding the presence of appendicitis and other findings related to appendicitis using syntactic dependency patterns. All relevant individual statements from a report are collectively considered to determine whether the report is consistent with appendicitis. Evaluation on a corpus of 400 radiology reports annotated by two expert radiologists showed that our approach achieves a precision of 91\%, a recall of 83\%, and an F1-measure of 87\%.

Introduction
The ability to automatically identify actionable findings in radiology reports, such as appendicitis can play an important role in clinical quality improvement research. As reported by the National Patient Safety Goals of the Joint Commission, up to 70\% of sentinel medical errors result from communication errors\textsuperscript{1}. Moreover, a natural language processing system may help alleviate the inadequate communication of critical results, which as reported by Huntington et al\textsuperscript{2} is the principal cause of malpractice cases involving radiologists in the USA.

In the research reported in this paper, we explore the use of natural language processing (NLP) techniques for identifying radiographic findings. We focus on a single actionable finding, namely the detection of records consistent with appendicitis. Appendicitis is a condition in which the appendix becomes inflamed. Our approach uses a method based on the recognition of patterns to detect direct assertions about appendicitis made by the radiologist along with common findings associated with appendicitis such as inflammation located in or around the appendix. These patterns use information about the syntactic structure of the text to determine anatomical location for mentions of inflammation. Syntactic structure is also used to correctly scope the usage of negation for mentions of appendicitis (e.g., “No findings of appendicitis.”) Our evaluation shows that this approach achieves an F1-measure of 85\%, with a precision of 84\%, and a recall of 87\% on a set of 400 radiology reports annotated by two radiologists.

Radiology corpus
We created a corpus of 7,230 de-identified radiology reports from the UT Southwestern Medical Center utCRIS data warehouse. The reports were then filtered to retrieve only those radiology reports possibly diagnostic for appendicitis. The corpus is divided into an unannotated development set of 6,830 reports, and an annotated test set of 400 reports. The test set consists of 100 reports which contain an indicator of inflammation (e.g., inflammation, thickening), 100 reports containing an anatomical term near the appendix (e.g., cecum), 100 reports containing both types of terms, and 100 reports containing none of these types of terms. The development set was examined for common language patterns in the reports. Those patterns were then encoded into the dependency patterns used in our approach for detecting appendicitis indicators in text. The testing set has 87 positive reports for indicating appendicitis and 313 negative reports.

Lexicons
Natural language, particularly within the domain of radiology reports, allows for a large degree of variation in expression. For example, some doctors will use the term “appendectomy” while others use “appendicectomy”. Furthermore, there may be similar problems, such as: inflammation, inflammatory changes, thickening, and dilatation. While not direct synonyms, for many purposes they can be treated as the same medical condition. When detecting appendicitis it is also important to know which anatomical structures are located near other anatomical structures. For instance, inflammation near the cecum is very indicative of appendicitis because the cecum is directly connected to the appendix. Lexicons are used to capture these variations in language. The two largest lexicons encode terms relating to inflammation or other abnormalities associated with appendicitis in different syntactic contexts. Excerpts from the inflammation and anatomy lexicons follow:

Nominal Inflammation Lexicon [INFLAMM]: abscess, appendicitis, appendicolith, change, changes, inflammation, mass, perforation, phlegmon, rupture, stranding, thickening
Adjectival and Verbal Inflammation Lexicon [INFLAMM_VERB_ADJ]: abnormal, change, dilated, distended, enlarged, inflamed, perforated, phlegmonous, ruptured, thickened

Anatomy lexicon [ANAT]: appendiceal, appendicitis, appendicolith, appendix, cecal, cecum, mesentery, periappendiceal, pericecal, periluminal, quadrant

The choice was made to create a lexicon limited to anatomical terms related to the appendix versus using a general vocabulary such as SNOMED CT\(^3\) or the Foundational Model of Anatomy\(^4\). Our examination of the radiology reports showed that a limited number of terms were used to describe the anatomy important to appendicitis. The lexicons listed in this section are referenced by the patterns used in our approach, which we describe in the next two sections.

**Identification of assertion status for mentions of appendicitis**

Often a radiology report will directly mention appendicitis in either an affirmative finding (*There are some strandy changes suggesting appendicitis*), or in a negative finding (*No evidence for appendicitis*).

When a radiology report mentions appendicitis, the default belief status is positive for appendicitis. Based on the context, though that belief status can be changed to negative. This is known to be a difficult problem as negation markers such as *no* and *not* need to be scoped to the correct term.\(^5,6\) We solve these problems by capturing the local syntactic context for a mention of appendicitis which is relevant to its assertion status. The context for a mention of appendicitis is composed of several components, including statements of belief and observational statements. Statements of belief are detected using phrases such as the following:

| Statements of belief | consistent with, concerning for, suggestive of, suspicious for, worrisome for, evidence of, possibility of, suspicion of, probability of |

These phrases are matched using dependency patterns such as [consistent prep with poly, appendicitis] and [compatible prep with poly, appendicitis]. These patterns allow the flexibility to match statements such as *worrisome for early appendicitis*. In addition to detecting belief statements, it is also necessary to detect observational cues such as *evidence of appendicitis* and *Examination indicates appendicitis*. Additional lexicons contain terms describing observation through the use of nouns and verbs.

The observation lexicons are:

| Nominal observation lexicon [OBSERV_NOUN]: finding, findings, examination, examinations, evidence, impression, impressions, appearance, appearances, pattern, patterns, diagnosis, signs, sign, etiology |

| Verbal observation lexicon [OBSERV_VERB]: seen, noted, identified, visualized, detected, found, observed |
Examples
Findings compatible with acute, nonperforated appendicitis.

Thickening of the tip of the cecum

ADJ VERB The appendix is mildly distended; patient's diseased appendix

VERB ADJ Acute appendicitis, potentially ruptured.

Dilated appendix with adjacent stranding

There is no periappendiceal inflammation

This may be perforated appendicitis.

There is no evidence of appendicitis

right lower quadrant mass

The previously dilated appendix is no longer dilated

Acute appendicitis, potentially ruptured.

Dilated appendix with adjacent stranding

right lower quadrant mass

Findings compatible with acute, nonperforated appendicitis.

An appendicitis is seen at the proximal appendix

| Pattern | Examples |
|---------|----------|
| INFLAMM \( \text{prep} \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | Inflammatory changes in the right lower quadrant |
| ANAT \( \text{amod} \rightarrow \text{INFLAMM\_VERB\_ADJ} \) | The appendix is mildly distended; patient’s diseased appendix |
| INFLAMM \( \text{amod} \rightarrow \text{ANAT} \) | There is no periappendiceal inflammation |
| INFLAMM\_VERB\_ADJ \( \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | The previously dilated appendix is no longer dilated |
| INFLAMM\_VERB\_ADJ \( \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | Thickening in the region of the cecum |
| INFLAMM\_VERB\_ADJ \( \rightarrow \text{nsubj} \rightarrow \text{ANAT} \) | This may be perforated appendicitis. |
| INFLAMM \( \text{prep} \rightarrow \text{with} \rightarrow \text{pobj} \rightarrow \text{INFLAMM} \) | Acute appendicitis, potentially ruptured. |
| OBSERV\_NOUN \( \text{prep} \rightarrow \text{of} \rightarrow \text{pobj} \rightarrow \text{appendicitis} \) | Dilated appendix with adjacent stranding |
| INFLAMM \( \text{prep} \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | right lower quadrant \( \rightarrow \text{INFLAMM} \) |
| OBSERV\_NOUN \( \text{nsubj} \rightarrow \text{consistent} \rightarrow \text{prep} \rightarrow \text{with} \rightarrow \text{pobj} \rightarrow \text{appendicitis} \) | right lower quadrant mass |
| INFLAMM \( \text{nsubj} \rightarrow \text{pobj} \rightarrow \text{OBSERV\_VERB} \rightarrow \text{prep} \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | Findings compatible with acute, nonperforated appendicitis. |

Table 1: The most frequently matched patterns. Words in ALL CAPS refer to lexicons. Literals are indicated in italics. <ROOT> refers to the root of the dependency parse tree. Words without any restrictions are denoted by [].

| Pattern | Examples |
|---------|----------|
| \( \text{INFLAMM} \rightarrow \text{prep} \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | Inflammatory changes in the right lower quadrant |
| \( \text{ANAT} \rightarrow \text{amod} \rightarrow \text{INFLAMM\_VERB\_ADJ} \) | The appendix is mildly distended; patient’s diseased appendix |
| \( \text{INFLAMM} \rightarrow \text{amod} \rightarrow \text{ANAT} \) | There is no periappendiceal inflammation |
| \( \text{INFLAMM\_VERB\_ADJ} \rightarrow \text{prep} \rightarrow \text{prep} \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | The previously dilated appendix is no longer dilated |
| \( \text{INFLAMM\_VERB\_ADJ} \rightarrow \text{prep} \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | Thickening in the region of the cecum |
| \( \text{INFLAMM\_VERB\_ADJ} \rightarrow \text{nsubj} \rightarrow \text{ANAT} \) | This may be perforated appendicitis. |
| \( \text{INFLAMM} \rightarrow \text{prep} \rightarrow \text{with} \rightarrow \text{pobj} \rightarrow \text{INFLAMM} \) | Acute appendicitis, potentially ruptured. |
| \( \text{OBSERV\_NOUN} \rightarrow \text{prep} \rightarrow \text{of} \rightarrow \text{pobj} \rightarrow \text{appendicitis} \) | Dilated appendix with adjacent stranding |
| \( \text{INFLAMM} \rightarrow \text{prep} \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | right lower quadrant mass |
| \( \text{OBSERV\_NOUN} \rightarrow \text{nsubj} \rightarrow \text{consistent} \rightarrow \text{prep} \rightarrow \text{with} \rightarrow \text{pobj} \rightarrow \text{appendicitis} \) | right lower quadrant mass |
| \( \text{INFLAMM} \rightarrow \text{nsubj} \rightarrow \text{pobj} \rightarrow \text{OBSERV\_VERB} \rightarrow \text{prep} \rightarrow \text{S-PREP} \rightarrow \text{pobj} \rightarrow \text{ANAT} \) | Findings compatible with acute, nonperforated appendicitis. |

Spatially locating inflammation

Although an explicit mention of appendicitis can be a strong indicator that the patient has the disease, many reports do not directly state such a finding because it is implied to an expert through other findings, e.g., An inflamed, perforated appendix is seen. Therefore some of the patterns detect detect mentions of inflammation and the specific anatomy associated with each mention. If the inflammation is associated with an anatomical structure near the appendix we consider that as evidence toward a finding of appendicitis.

The first set of patterns for spatially locating inflammation relies on a prepositional relation between a mention of inflammation and an anatomical term such as: inflammation around the appendix or appendix with inflammation. The set of prepositions considered for spatial relations is:

**Spatial prepositions [S-PREP]:** around, about, of, to, near, at, within, into, in, throughout, along

Anatomical component lexicon: wall, walls, tip, base, junction, valve, lumen, segment, segments, region, area, aspect, location, fat, orifice, opening, entrance, mucosa, apex, mass, masses, plane, planes

A component is associated with its larger structure through the patterns: [component] of [structure], [structure-adjective] [component], junction with [structure]. These patterns allow for matching examples such as cecal tip thickening or stranding near the junction with the appendix.

Classification

The indicators for appendicitis we detect in radiology reports have varying degrees of effectiveness for predicting whether the patient has appendicitis. For example, if we detect Patient has acute appendicitis we are much more confident in an assessment of appendicitis than if we only saw cecum is inflamed. However, the latter statement does provide some evidence towards a decision of appendicitis. Also, the more evidence that we see for a decision of appendicitis, the more confidence we can give to our decision. We use an SVM classifier to make a decision about whether a report indicates appendicitis. The features for the classifier are all based on the indicators we find in the report. The simplest type of features capture individual words that have been detected in the report along with their semantic role (e.g. anatomy= appendix, inflammation=thickening). We also have features which capture the combination of anatomy and inflammation together (e.g. (inflammation, cecum)). The final class of features capture all roles matched in a pattern along with their content (e.g. (observation= evidence, inflammation=inflamed, anatomy= appendix)). All of the features also capture whether the indicator was negated or not.
While our approach performs well, it does still make some errors in judging reports for indications of appendicitis. Given the many components comprising the approach, it is informative to analyze why errors are made. One source of errors arises from the lack of section identification. Reports which mention the word appendicitis are marked as indicative of appendicitis. Therefore an approach which marked all reports as non-indicative of appendicitis would achieve an accuracy of 78.3%, while our method achieved an accuracy of 93.8%, a significantly better result. We also evaluated an approach which only marked documents as indicating appendicitis if the word appendix appeared in the report. The results show that 74.71% of reports indicative for appendicitis mention the word appendicitis. The F1 measure of such a system is a full nine points lower than our approach. The patterns added by the automatic pattern discovery method contributed to a gain in F1 of almost 1.6 points.

### Error Analysis

While our approach performs well, it does still make some errors in judging reports for indications of appendicitis. Given the many components comprising the approach, it is informative to analyze why errors are made. One source of errors arises from the lack of section identification. Reports which mention *evaluate for appendicitis* in the history section, but which later indicate a normal appendix are marked as indicative of appendicitis.

### Results

Our test set consists of 400 radiology reports from UT Southwestern Medical Center. The authors ST and TB annotated the reports as positive or negative for appendicitis. Using our lexicons for anatomical terms near the appendix and inflammation terms, we created four groups of 100 reports each: (1) reports containing at least one term from both lexicons, (2) reports containing only one term from the appendicitis lexicon, (3) reports containing only one term from the anatomical lexicon, and (4) reports containing neither inflammation nor anatomical terms. This resulted in a more balanced training set, since many of the initial reports were not relevant to appendicitis.

### Evaluation

We evaluated our approach using the F1 measure. The F1 measure is a metric for evaluating both the precision and recall of a classification method. The formula is $F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$. Table 2 shows the results of our evaluation. The evaluation was performed using a 5-fold cross validation. The testing set has 87 positive reports and 313 reports negative for indicating appendicitis. Therefore an approach which marked all reports as non-indicative of appendicitis would achieve an accuracy of only 78.3%, while our method achieved an accuracy of 93.8%, a significantly better result. We also evaluated an approach which only marked documents as indicating appendicitis if the word appendicitis appeared in the report. The results show that 74.71% of reports indicative for appendicitis mention the word appendicitis. The F1 measure of such a system is a full nine points lower than our approach. The patterns added by the automatic pattern discovery method contributed to a gain in F1 of almost 1.6 points.
Several more rounds of interactions with the doctors would be necessary to fine-tune our lexicons and patterns. For example, *contrast* is somewhat erroneously included as a term of inflammation. There are two types of contrast often given for a CT scan. Oral contrast to define the lumen of the stomach and intestines and IV contrast which opacifies the blood. If there is "contrast enhancement" or just "enhancement" of a given area it is a sign that there is increased blood flow to that region possible caused by an infection or inflammatory process. The term *contrast* is often used in contexts such as: *The appendix is filled with oral contrast and there is no evidence of small or large bowel obstruction.* In this context *contrast* should be understood to mean that oral contrast was able to freely enter and fill the appendix and hence there is nothing blocking flow into the appendix.

**Conclusion**

In this paper we described an approach for identifying radiology reports which indicate appendicitis. The approach is based on recognizing direct assertions of appendicitis and indirect evidence such as indications of inflammation near the appendix. These indications are detected through the use of patterns which capture the syntactic dependency structure of the text. Various linguistic phenomenon such as negation and conjunctions are handled. Each report is automatically categorized as indicative or non-indicative of appendicitis by combining all relevant statements found within the reports. Direct indications of appendicitis are given more influence in the decision and indications which are general or specific to anatomical structures other than the appendix are given less influence. Our evaluation shows that our approach can identify reports consistent with appendicitis with a precision of 84%, a recall of 87% and an F1 score of 85. Error analysis has revealed promising directions for relatively easy improvement including section identification and continued input from the doctors.

The automatic identification of actionable findings in radiology reports can lead to improvements in patient outcomes and quality of care. One application could be an interpretive software layer which could alert the referring clinician about radiology reports containing language associated with a given disease process. A second application would be enabling the radiologist to ensure that a given outcome resulted from the language they used.

**References**

1. Lucey L.L. and Kushner D.C. The acr guideline on communication: to be or not to be, that is the question. *Journal of the American College of Radiology*, 7(2):109–114, 2010.
2. Huntington B. and Kuhn N. Communication gaffes: a root cause of malpractice claims. *Proceedings (Baylor University. Medical Center)*, 16(2):157, 2003.
3. Stearns M.Q., Price C., Spackman K.A. and Wang A.Y. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662, 2001.
4. Rosse C., Mejino J.L.V. and others. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500, 2003.
5. Huang Y. and Lowe H.J. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304–311, 2007.
6. Apostolova E., Tomuro N. and Demner-Fushman D. Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes. In *Annual Meeting of the ACL*, pages 283–287, 2011.
7. Lassila O. and Swick R.R. Resource description framework (rdf) model and syntax. *World Wide Web Consortium*, http://www.w3.org/TR/WD-rdf-syntax, 1999.
8. Prud’Hommeaux E. and Seaborne A. SPARQL query language for rdf. *W3C working draft*, 4(January), 2008.