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

Bryan Rink[1], Kirk Roberts[1], Sanda Harabagiu, PhD[2], Richard H. Scheuermann, PhD[2], Seth Toomay, MD[2], Travis Browning, MD[2], Teresa Bosler, PMP[2], Ronald Peshock, MD[2]
[1] The University of Texas at Dallas, Richardson, TX, USA
[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 [1]. Moreover, a natural language processing system may help alleviate the inadequate communication of critical results, which as reported by Huntington et al [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*

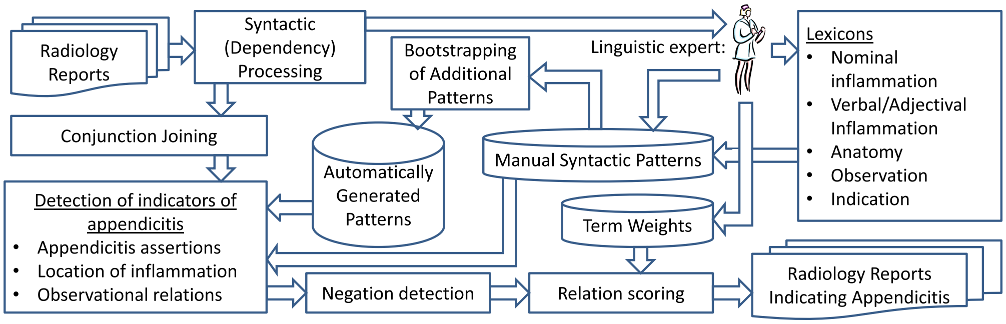
Radiology Reports
Syntactic (Dependency) Processing
Bootstrapping of Additional Patterns
Linguistic expert:
Lexicons
Nominal inflammation
Verbal/Adjectival Inflammation
Anatomy
Observation
Indication
Conjunction Joining
Automatically Generated Patterns
Manual Syntactic Patterns
Detection of indicators of appendicitis
Appendicitis assertions
Location of inflammation
Observational relations
Term Weights
Negation detection
Relation scoring
Radiology Reports Indicating Appendicitis

Figure 1: Architecture of the framework for identifying indications of appendicitis in radiology reports.

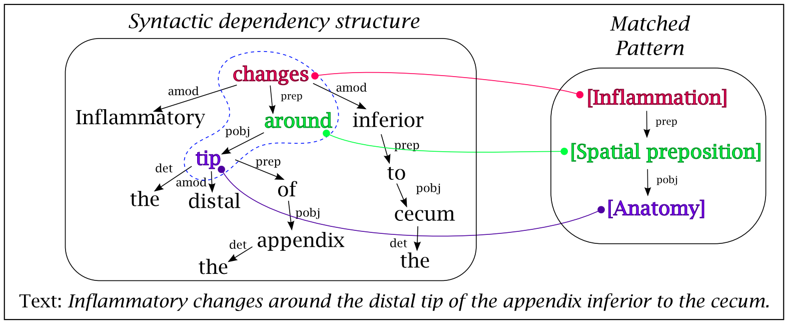

Figure 2: The syntactic structure of a typical statement indicative of appendicitis and a pattern which matches it.

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 $\xrightarrow{prep}$with $\xrightarrow{pobj}$ appendicitis] and [compatible $\xrightarrow{prep}$ with $\xrightarrow{pobj}$ 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 is:

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*

Pattern	Examples
INFLAMM $\xrightarrow{prep}$ S-PREP $\xrightarrow{pobj}$ ANAT	Inflammatory changes in the right lower quadrant
ANAT $\xrightarrow{amod}$ INFLAMM_VERB_ADJ	The appendix is mildly distended; patient's diseased appendix
INFLAMM $\xrightarrow{amod}$ ANAT	There is no periappendiceal inflammation
INFLAMM_VERB_ADJ $\xrightarrow{nsubjpass}$ ANAT	The previously dilated appendix is no longer dilated
INFLAMM_VERB_ADJ $\xrightarrow{prep}$ S-PREP $\xrightarrow{pobj}$ ANAT <ROOT> $\rightarrow$ *appendicitis*	Thickening in the region of the cecum This may be perforated appendicitis.
INFLAMM_VERB_ADJ $\xrightarrow{nsubj}$ ANAT	Acute appendicitis, potentially ruptured.
ANAT $\xrightarrow{prep}$ *with* $\xrightarrow{pobj}$INFLAMM	Dilated appendix with adjacent stranding
OBSERV_NOUN $\xrightarrow{prep}$ *of* $\xrightarrow{pobj}$ *appendicitis*	there is no evidence of appendicitis
INFLAMM $\xrightarrow{prep}$ S-PREP $\xrightarrow{pobj}$ [] $\xrightarrow{prep}$ S-PREP $\xrightarrow{pobj}$ ANAT	thickening of the tip of the cecum
right lower quadrant $\xleftarrow{nn}$ INFLAMM	right lower quadrant mass
OBSERV_NOUN $\xrightarrow{nsubj}$ *consistent* $\xrightarrow{prep}$ *with* $\xrightarrow{pobj}$ *appendicitis*	Findings compatible with acute, nonperforated appendicitis.
INFLAMM $\xleftarrow{nsubjpass}$ OBSERV_VERB $\xrightarrow{prep}$ S-PREP $\xrightarrow{pobj}$ ANAT	An appendicolith is seen at the proximal appendix

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 [].

Indication lexicon: *indicate, indicating, indicates, represent, representing, represents, suggest, suggesting, suggests, reflect, reflecting, reflects, shows, show, demonstrates, demonstrate, presents, present*

These expressions are combined in multiple ways to form dependency patterns, as shown in Table 1. In addition to the patterns for positive assertions, we have a small handful of negative patterns as well including [excludes $\xrightarrow{nsubj}$appendicitis] and [excluded $\xrightarrow{nsubjpass}$ appendicitis] (e.g., *appendicitis is excluded*). Matches for negative patterns are checked for negation as well, and invert the assertion if they are negated. For example *appendicitis is not excluded* would be considered a positive assertion.

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*

Sometimes the radiologist will specify that inflammation has occurred within a specific component of an anatomical structure, such as a wall or a tip, as in *inflamed appendiceal tip.* A lexicon exists for detecting components of anatomical structures so that spatial relation detection can be performed between inflammation and a component, which is then associated with a structure such as *appendix* or *colon*.

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.

Method	Accuracy	Precision	Recall	F1
Always “Non-indicative”	78.3%			
“Appendicitis” only	90.5%	78.31	74.71	76.47
Manual patterns only	93.3%	83.52	87.36	85.39
Manual and automatically generated patterns	94.3%	91.14	82.76	86.75

Table 2: Evaluation for the detection of indications of appendicitis in radiology reports.

Automatic Pattern Discovery

Depending upon a set of manually crafted rules presents two primary limitations: (1) radiologists express their findings using a diverse set of lexical and syntactic expressions, and (2) manually created patterns are generally expressed assuming proper automatic recognition of linguistic structure, whereas syntactic parsers make many mistakes on clinical text. To overcome these limitations, we propose a method for automatically generating syntactic dependency patterns to increase the recall of actionable finding extraction. Our method leverages our manually created patterns in combination with dependency transformation operators to discover new patterns.

Two types of transformation operators are applied to the manually created patterns to automatically create candidate patterns. The first operator replaces lexicon and literal restrictions on a single token with a wildcard to match all possible words. For example, given the pattern (“[OBSERV_NOUN] $\xrightarrow{prep}$of $\xrightarrow{pobj}$ *appendicitis*”, it would create three patterns: (1) “* $\xrightarrow{prep}$ of $\xrightarrow{pobj}$ *appendicitis*”, (2) “[OBSERV_NOUN] $\xrightarrow{prep}$ * $\xrightarrow{pobj}$ *appendicitis*”, and (3) “[OBSERV_NOUN] $\xrightarrow{prep}$ of $\xrightarrow{pobj}$ *”. The second operator replaces each dependency relation with a wildcard. For the above pattern, the expansions would be: (4) “[OBSERV_NOUN] $\xrightarrow{*}$ of $\xrightarrow{pobj}$ *appendicitis*” and (5) “[OBSERV_NOUN] $\xrightarrow{prep}$ of $\xrightarrow{*}$ *appendicitis*” Next, we search our dependency-parsed corpus of radiology reports for matches to these patterns. For each match, we observe the word or dependency relation which corresponds to the wildcard (e.g., the wildcard for pattern (1) above might match the word “indication”). We refer to a specific word or dependency relation matching a wildcard as a grounding. All grounded candidate patterns which occur less than three times in the corpus are discarded, while the remainder are kept for scoring. Wildcard matching and occurrence counting for grounded patterns were implemented by representing the dependency parses for sentences using RDF[7] and storing them in a Sesame (openrdf.org) native RDF store. Sentences matching candidate patterns were then retrieved by issuing the appropriate database queries (using SPARQL[8]).

Each candidate pattern is assigned a score using Fisher’s exact test comparing the documents matched by a pattern and the documents marked as positive for appendicitis by the manually created patterns. Candidate patterns which match many of the reports marked as positive and few of the documents marked as negative will be assigned the highest scores. The top ranking patterns are then examined by a human to prune out patterns that would not be expected to generalize well on a new set of reports. Using this technique we were able to add 39 patterns positive for appendicitis and 3 patterns negative for 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 an inflammation term, but not an anatomical term, (3) reports containing an anatomical term, but not an inflammation term, 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 F_1 measure. The F_1 measure is a metric for evaluating both the precision and recall of a classification method. The formula is $F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$. 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.

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.

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 an 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.