Automatic Identification of Critical Follow-Up Recommendation Sentences in Radiology Reports

Journal:	*AMIA 2011 Annual Symposium*
Manuscript ID:	AMIA-0619-A2011.R1
Manuscript Type:	Paper

Automatic Identification of Critical Follow-Up Recommendation Sentences in Radiology Reports

Meliha Yetisgen-Yildiz, PhD[1,3], Martin L. Gunn, MB ChB, FRANZCR[2], Fei Xia, PhD[3,1], Thomas H. Payne, MD[4,1]

[1]Biomedical & Health Informatics, School of Medicine, University of Washington, Seattle, WA
[2]Department of Radiology, School of Medicine, University of Washington, Seattle, WA
[3]Department of Linguistics, University of Washington, Seattle, WA
[4]Information Technology Services, School of Medicine, University of Washington, Seattle, WA

Abstract

Communication of follow-up recommendations when abnormalities are identified on imaging studies is prone to error. When recommendations are not systematically identified and promptly communicated to referrers, poor patient outcomes can result. Using information technology can improve communication and improve patient safety. In this paper, we describe a text processing approach that uses natural language processing (NLP) and supervised text classification methods to automatically identify critical recommendation sentences in radiology reports. To increase the classification performance we enhanced the simple unigram token representation approach with lexical, semantic, knowledge-base, and structural features. We tested different combinations of those features with the Maximum Entropy (MaxEnt) classification algorithm. Classifiers were trained and tested with a gold standard corpus annotated by a domain expert. We applied 5-fold cross validation and our best performing classifier achieved 95.60% precision, 79.82% recall, 87.0% F-score, and 99.59% classification accuracy in identifying the critical recommendation sentences in radiology reports.

Introduction

Most medical reports are created and stored in natural language. Radiology reports, which are a description of relevant disease processes found by radiologists on imaging studies, such as radiographs, computed tomography (CT) scans, are one such class of medical documents and contain a rich semantic vocabulary designed to interpret imaging data with text. If a radiologist makes a potentially important observation when examining an imaging study he/she may make further specific recommendations for follow-up imaging tests, or clinical follow-up in his/her report. These recommendations are made when the radiologist considers the finding to be clinically significant, and unexpected, and believes that it is important for the referring physician to consider further investigation, management, or follow-up of the finding in order to avoid an adverse outcome. Moreover, radiologists may make clinically important or unexpected findings in imaging studies that need to be verbally communicated with the physician caring for the patient. Although there is no precise standard definition for this type of observation at present, these findings are often termed *critical results*. Similar to Hussain, we consider a critical result to be finding for which reporting delays can result in serious outcomes for patients.[1] The American College of Radiology (ACR) recommends that radiologists supplement their written report with “non-routine” means of communication with the treating or referring physician (usually verbal) to ensure adequate receipt of the information in a timely manner.[2] The National Patient Safety Goals of the Joint Commission recommend verbal communication of these results, with documentation of the communication. For example, the finding of a lung nodule suspicious for cancer in a patient undergoing a CT scan following a motor vehicle collision would be classified as a critical result, and should be verbally communicated with the referrer. An important related problem may occur when verbal communication occurs with the ordering clinician, but another clinician assumes responsibility for the patient later, and ”hand-off” of what may seem at the time to be a non-urgent medical problem does not occur. Other proposed safety net procedures include communication acknowledgement systems, critical result work lists, and compliance audit systems. [3]

Despite the imperative of good communication to avoid medical errors, it does not always occur. Inadequate communication of critical results is the cause of the majority of malpractice cases involving radiologists in the USA.[4] The Joint Commission reported that up to 70% of sentinel medical errors were caused by communication errors.[5] The Institute of Medicine reported in its document, *To Err Is Human: Building a Safer Health System*,[6] that each year around 98,000 patients die due to potentially preventable errors and has given directions to improve safety. In that report, the emphasis on communication errors reflects both a long standing directive for the medical community as well as an increasing recognition that delays, failures, and errors in the communication of important test results can and do threaten patient safety.

The motivation for the research study presented in this paper was our observation that sometimes radiologists themselves do not recognize that their own report contains an important recommendation for further clinical or imaging follow-up, and that radiologists do not instigate verbal communication in these settings, or identify them as critical and the recommendation may be included only in the narrative text body of the radiology report, and not in the “Impression” or ”Summary” section, and thus not highlighted. Hence these potentially important observations and recommendations might not be apparent to clinicians caring for the patient. In large institutions, the clinician who ends up following the patient is not the same as the person who ordered the imaging study. Even if the need to consider future investigation was verbally communicated to the ordering clinician, this information may be missed several months later when someone else views a long list of reports in the EMR. A recommendation will not be considered by the treating clinician if it is not seen. The goal of our research is to identify important follow-up recommendations (most commonly these are imaging tests, but they also include relevant blood tests, endoscopy or eliciting specific clinical signs) so that the reports can be flagged and separate workflow processes can be initiated to reduce the chance that needed investigations suggested in the report are missed by clinicians, and as a result, further action not considered. As an initial step to accomplish this goal, we designed a text processing approach to identify the sentences that involve critical recommendation information. In this research study, we defined *critical recommendation* as a statement made by the radiologist in a given radiology report to advise the referring clinician to further evaluate an imaging finding by either other tests or further imaging. In the remaining of this paper, for the sake of simplicity, we use *recommendation* to refer to *critical recommendation* unless specified otherwise.

To identify recommendation sentences, we applied natural language processing (NLP) and supervised text classification methods to free text available in radiology reports. To increase the identification performance, we enhanced the simple bag-of-words text representation approach with lexical and semantic features. We evaluated the performance by comparing the system predictions against a gold standard created by a domain expert. Based on the good performance of the NLP system, we plan to incorporate this system to the current production clinical computing systems used in UW Medicine to reduce medical errors caused by miscommunications between radiologists and clinicians.

Related Work

In the clinical NLP domain, radiology reports have been widely studied by various reseachers.[7-12] As one of the earliest examples, Friedman et al. developed and evaluated a text processor called MedLEE (Medical Language Extraction and Encoding System) that extracts and structures clinical information from textual radiology reports and translates the information to terms in a controlled vocabulary so that the clinical information can be accessed by further automated procedures.[7,8] Jain et al. used MedLEE to encode the clinical information in chest and mammogram reports to identify suspected tuberculosis[9] and breast cancer[10] patients. Hersh et al. described an NLP system called SAPHIRE that matched text to concepts in the Unified Medical Language System (UMLS) metathesaurus for automatic indexing of radiology reports to develop clinical image repositories that can be used for patient care and medical education.[12] In this paper, our overall goal was to identify sentences that include critical recommendation information in radiology reports.

The problem of identification of recommendation information in radiology reports has also been previously studied by other researchers.[13-15] Dang et al. processed 1059 radiology reports with Lexicon Mediated Entropy Reduction (LEXIMER) to identify the reports that include clinically important findings and recommendations for subsequent action.[13] In that study the researchers did not analyze the documents in the sentence level. The same research group performed a similar analysis on a database of radiology reports covering the years 1995-2004.[14] From that database, they randomly selected 120 reports with and without recommendations. Two radiologists independently classified those selected reports according to the presence of recommendation, time-frame, and imaging-technique suggested for follow-up examination. These reports were analyzed by an NLP system first for classification into two categories: reports with recommendations and reports without recommendations. The reports with recommendations then were classified into those with imaging recommendations and those with non-imaging recommendations. The recommended time frames were identified and normalized into number of days. The authors reported 100% accuracy in identifying reports with and without recommendations. In 88 reports with recommendation, they reported 94.5% precision in identifying temporal phrases, and 93.2% in identifying recommended imaging tests. In a follow-up study, the authors analyzed the rate of recommendations by performing a statistical analysis on 5.9 million examinations[15]. In all three papers, they reported impressive overall performance values; however, the authors presented their text processing approach as a black box without providing necessary information required to replicate their methods.

In this paper, our main research focus was on investigation of different text representation approaches to increase the performance of recommendation sentence classification in radiology reports. To accomplish this, we built a text processing approach and described its main steps in detail to ensure that our methods can be replicated and evaluated by other researchers.

Methods

The overall architecture of our text processing approach for recommendation sentence extraction can be found in Figure 1. In the following sections, we will explain the main steps of our text processing approach in detail.

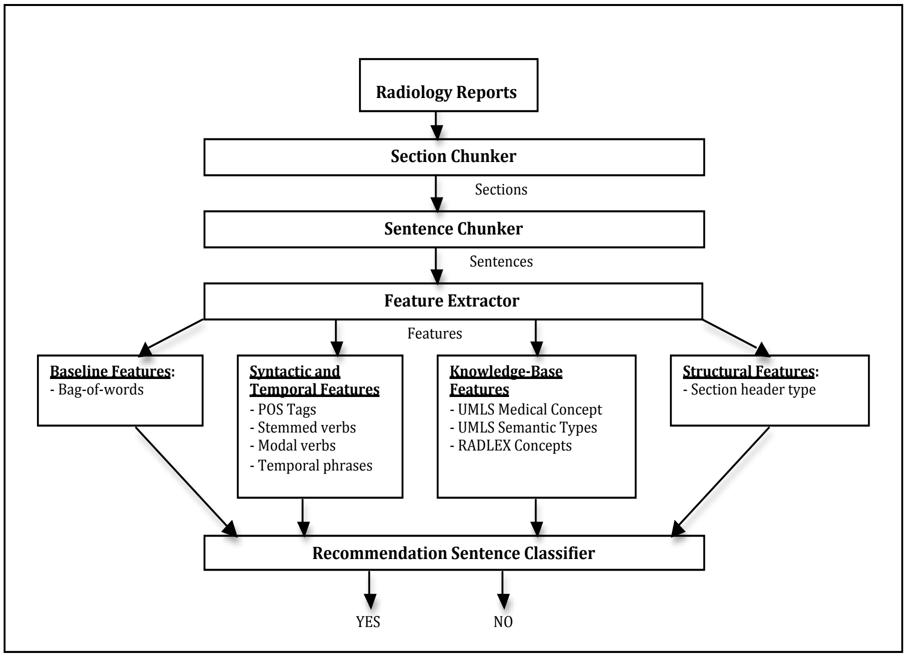

Figure 1. System Architecture

Radiology Corpus

We created a corpus of radiology reports composed of 800 de-identified radiology reports extracted from Harborview Medical Center Radiology Information System. The reports had been generated using the Nuance RadWhere™ radiology voice recognition reporting system using a combination of structured report templates, speech recognition, and freely typed text. The reports were stored on General Electric Centricity Radiology Information System (RIS) version 10.6. The reports represented a mixture of imaging modalities, including radiographs, CT scans, ultrasounds and magnetic resonance imaging (MRI). The retrospective review of those reports was approved by the University of Washington Human Subjects Committee of Institutional Review Board, who waived the need for informed consent. The second author of the paper (M.L.G.)[a], who is a radiologist, examined each report and annotated the sentences that include a recommendation. In the final gold standard, 81 of 800 reports (10.12%) included recommendations and a total of 109 sentences out of 6378 sentences in the 800 reports were annotated as recommendation sentences.

Section Chunker

Although radiology reports are free text, they are somewhat structured in terms of sections. Friedman categorized those sections under three general headings: *Clinical Information*, which usually contains some information related to the indication of the examination, *Description* which describes the findings, and *Impression* which lists or interprets the most relevant findings.[16]

[a] M.L.G. participated in human annotation and error analysis phases of the study. He did not actively participated in the system design and development phases to prevent introducing bias to the system performance.

To understand the section characteristics of our radiology corpus, we randomly selected 50 reports and manually annotated the section headers. In total, including repeated headers, we identified a total of 297 section headers in those 50 reports. 54 of those headers were main report titles such as “CHEST RADIOGRAPHS, TWO VIEWS”. We classified the remaining 243 headers into Friedman's three categories. The distribution of the headers is presented in Table 1.

Section Category	Section Header	Frequency
Clinical Information	clinical history	4
	clinical indication	1
	comparison	44
	contrast	12
	exam	3
	history	22
	indication	26
	procedure	17
	protocol	1
	technique	19
Description	findings	47
	preliminary findings	3
Impression	attending comment	2
	critical result	1
	final attending report	1
	impression	39
	note	1

Table 1. The distribution of Section headers (243 headers from 50 reports) over three main section categories.

We also analyzed the style characteristics of the annotated headers and identified the following characteristics.

- 76.1% (226/297) of the headers ended with ":" (e.g., “CLINICAL HISTORY:”)
- 88.9% (264/297) of the headers were in all upper-case (e.g., “IMPRESSION:”)
- 11.1% (33/297) of the headers stated with a capital letter and continued with lower-case letters (e.g., “Comparison:”)

We designed regular expressions to capture the listed style characteristics and used those regular expressions to identify the lines that potentially include the section headers. Next, we manually wrote rules based on the frequent words that appear in the annotated headers listed in Table 1 to identify the correct headers and their general categories. For example, a given potential header string included the term *finding* should be labeled as a Description section header. Because the vocabulary used in the headers was quite limited, with a small set of rules, we managed to identify the section headers in a very accurate way.

In our corpus of 800 reports, our approach identified a total of 4703 section headers (clinical information: 3150, description: 866 and impression: 687). Many of those 4703 headers appeared in multiple reports and the number of unique headers was 104 (e.g., 467 reports had a section titled “*Impression*:”). We labeled the text chunks between the identified section headers as section bodies and assigned them to the corresponding section headers.

Sentence Chunker

After identifying the report sections, we used National Centre for Text Mining (NaCTeM) sentence chunker[b] to identify the boundaries of the sentences in the section bodies. We identified 6378 sentences (109 positive for recommendation, 6269 negative) in the 800 reports. The details of the identified sentences' distribution over the general section categories are available in Table 2.

Feature Extractor

To identify recommendation sentences, the first step is to find the features that capture the characteristics and content of sentences in an effective way. Feature representation has a direct affect on the overall classification performance and we used the following four feature types in our classification task.

[b] NacTeM Sentence and Paragraph Breaker. Available at: http://www.nactem.ac.uk/software.php

Sentence Class Type	Section Category	Frequency
Recommendation Sentence	Clinical Information	1
	Description	15
	Impression	93
Non-Recommendation Sentence	Clinical Information	1135
	Description	3836
	Impression	1298

Table 2. Recommendation and non-recommendation sentence distribution over three main section categories.

1) *Baseline Features*

Information retrieval research suggests that words work well as representation units for retrieving documents.[17] In the bag-of-words representation, each distinct word corresponds to a feature with a weight as its value that is correlated to the number of times the word occurs in the document. We observed that words usually do not appear more than once in a sentence and used a binary weighting approach in our representation. The vector representation of a sentence had 1 as the weight of a feature if the feature appeared in the sentence and 0 otherwise.

2) *Syntactic and Temporal Features*

We defined the following features to capture the syntactic and temporal characteristics of the radiology report sentences.

Part-Of-Speech (POS) tags: We extended the baseline bag-of-words representation by linking the POS tag to each word that appeared in the sentence. We used Stanford POS Tagger[18] to identify the POS tags of the words in sentences and attached the identified POS tag information to each word and used it as a feature (e.g., *recommended* (baseline feature) ⇒ *recommened_VBN*).

Bigrams: We used word bi-gram features to capture interesting multi-word features.

Tense: We defined a feature from the POS tags of the verbs in a given sentence to detect any patterns related to the tense of the verbs.

Modal verbs: A large set of positive sentences included modal verbs (e.g., "Given the small size, this lesion could be followed up in 6 months."). The percentage of negative sentences with modal verbs was much lower (158/6269=2.52%) than that of positive sentences (43/109=39.45%). We captured this characteristic by using a binary feature called *includesModalVerb* in our representation. For a given sentence, *includesModalVerb* feature was set to 1 if the sentence includes a modal verb and 0 otherwise.

Stemmed verbs: The most common verbs in the positive sentences are *recommend*, *suggest*, *consider*, and *advise*. However, the verbs can take multiple forms. As an example, the verb *recommend* was used as *recommend* 12 times and as *recommended* 22 times in the positive sentences. To prevent feature splits like this, we stemmed verbs with the Porter Stemmer[19] (e.g., *recommended* ⇒ *recommend*) and used those stemmed verbs as binary features in the representation.

Temporal phrases: Another important characteristic of recommendation sentences was that they occasionally included a temporal phrase to indicate the timing of the recommended event (e.g., "Given the small size, this lesion could be followed up in *6 months*"). To capture this characteristic, we implemented a rule-based temporal phrase extractor. We defined a lexicon that included words (e.g., day, month, today, and year) and numeric values used in temporal phrases. Based on that lexicon, we defined a rule set to identify temporal phrases (e.g., *in 6 months*). When we applied to our corpus, we observed that 31.19% (34/109) of positive sentences and 6.01% (377/6269) of negative sentences included temporal phrases. We defined a new binary feature called *includesTemporalPhrase* and set it to 1 if a given sentence includes a temporal phrase and to 0 otherwise.

3) *Knowledge-Base Features*

We used the following three knowledge-base features to capture the semantics of the radiology sentences.

UMLS Concepts: Representing the content with bag-of-words approach has two challenges. First, the percentage of multi-word phrases, such as *head injury*, in medical vocabulary is very high. This high prevalence of phrases represents a problem for text classification. For example, the meaning of *head injury* is very different from that of *head* alone. Second, synonymy is a very common characteristic among the medical phrases. For example, radiologists use *liver* and *hepatic* interchangeably even in the same reports. If not grouped explicitly, synonymous words or phrases are represented as different features in the feature vector, which leads to two major drawbacks. The first drawback is that the increase in the dimensionality of feature space is known to have a negative effect on the classification performance. The second drawback is that information is lost due to feature splits. Instead of having a stronger feature, the representation has multiple relatively weaker features that are synonyms of each other.

To identify biomedical phrases, our system uses a knowledge based, natural language processing approach to process the radiology report sentence. A key part of our approach is to use a knowledge base to help to identify the domain specific terms. The biomedical domain already has a large publicly available knowledge base called the Unified Medical Language System (UMLS)[c]. In the latest version of UMLS, there are over 2.3 million biomedical concepts as well as over 8.5 million concept names. To identify the biomedical phrases, we used MetaMap, a tool created by NLM that maps the strings in free text to biomedical concepts in the UMLS[d]. MetaMap uses the UMLS to find the closest matching known concept to each identified phrase in the free text. Our system sends each sentence to MetaMap and uses the Concept Unique Identifier (CUI) of the identified UMLS concepts to group the synonymous concepts. We used the identified CUIs as binary feature in our representation.

UMLS Semantic Types: Another knowledge source available in UMLS is the Semantic Network. Semantic Network is a directed graph composed of 135 categories called *semantic types* and 49 different relations defined between the semantic types. Each medical concept in UMLS is mapped to at least one semantic type.

Recommendation sentences often mention a radiology test. As an example, the sentence "While these may represent hemorrhagic cysts, further evaluation with *renal ultrasound* could be obtained to exclude renal neoplasm" recommends an ultrasound for further evaluation. The radiology tests are categorized under the UMLS Semantic Type *Diagnostic Procedure*. To capture this characteristic of recommendation sentences, for each sentence, we used the semantic types associated with the UMLS concepts, which are identified by MetaMap, as binary features in our semantic representation.

RadLex: The UMLS Semantic Type Diagnostic Procedure includes more than 37 thousand medical concepts. The number of tests mentioned in the recommendation sentences is only a tiny fraction of this amount. To make the semantic representation more specific to radiology tests, we used RadLex, which was developed under the leadership of the Radiological Society of North America to create a terminology that can be used to annotate, index, and retrieve content from the Medical Imaging Resource Center (MIRC).[20] RadLex is not a part of UMLS; however, it is made publicly available by RSNA[e]. To be able to link the medical concepts identified by MetaMap with Radlex, we queried each RadLex concept against UMLS and if there is match we assigned the corresponding CUIs in UMLS to the queried RadLex concept. There were 11,962 concepts in RadLex and 8,652 of them were in UMLS. To identify the radiology specific tests, we selected the RadLex concepts in UMLS with the semantic type Diagnostic Procedure. We defined a binary feature called *includesRADLEXConcept* for each sentence and set it to 1 if the sentence includes at least one RadLex radiology test concept and to 0 otherwise.

4) *Structural Features*

Recommendation sentences usually appear in *Impression* sections. As can be seen in Table 2, 85.32% (93/109) of the positive sentences appeared in the *Impression* sections. The remaining 13.76% (15/109) was in the *Description* sections and 0.9% (1/109) was in the *Clinical Information* sections. Such a distribution indicated the potential importance of section header information in the classification decision. In our approach, each sentence was assigned to one of the three section header categories and we used that header information as a binary feature in our representation.

Recommendation Sentence Classifier

After representing the content in radiology report sentences with syntactic, temporal, knowledge-base and structural features, we trained classifiers to identify the recommendation sentences. For our classification task, we picked the Maximum Entropy (MaxEnt)[21] algorithm due to its good performance in text classification tasks. In our experiments, we used the MaxEnt implementation in a machine learning package called Mallet[f].

Results

In our experiments, as the gold standard, we used the corpus composed of 800 radiology reports and annotated by the second author of the paper (M.L.G).

[c] Unified Medical Language System Fact Sheet. Available at: http://www.nlm.nih.gov/pubs/factsheets/umls.html

[d] Metamap (MMTx). Available at:http://mmtx.nlm.nih.gov/

[e] Radlex. Available at: http://www.rsna.org/radlex/downloads.cfm

[f] Mallet. Available at: http://mallet.cs.umass.edu

Metrics

We evaluated the classification performance by using precision, recall, F1, and accuracy performance metrics. Because there was a limited number of positive sentences in our annotated corpus (109/6378=1.7%), we decided to use 5-fold cross validation to measure the performance of our classifiers.

Classification Performance

We measured the classification performance of MaxEnt, with four different feature types (baseline, syntactic & temporal, knowledge-base, and structural features). The number of distinct features introduced to the representation by each feature type is given in Table 3.

Feature Type	Feature Sub-type	# of Distinct Features
Baseline (B)	Word unigram	3494
Syntactic & Temporal (S)	POS	3547
	Word bigram	16905
	Tense	6
	Stemmed Verb	279
	includesModalVerb	1
	includesTemporalPhrase	1
Knowledge-base (K)	UMLS Concept	3445
	UMLS Semantic Type	104
	includesRADLEXConcept	1
Structural (St)	Section Type	3

Table 3. Feature set sizes. B: Baseline features, S: Syntactic and temporal features, K: Knowledge-base features, St: Structural features.

To understand the effect of each feature type, we added knowledge-base and structural features to the baseline features individually and compared the classification performance. We didn't combine the baseline features with the syntactic & temporal features because in syntactic features we extended the baseline bag-of-words approach with POS tags by attaching the POS tag information to each word (e.g., *recommended* ⇒ *recommended_VBN*).

To see the effect of all features, we combined syntactic & temporal features, knowledge-base features, and structural features as the last feature combination, and compared its classification performance with the baseline performance. Table 4 includes the classification performance for the baseline features and the four different feature combinations described above. The accuracy was high for all experiment since most instances were negative; therefore precision, recall, and F-score were more informative for evaluation purposes. As can be seen from Table 4, MaxEnt achieved the best recall and F-score values with the combined features (S+K+St) and the best precision value with the Syntactic and Temporal features (S).

Feature Types	MaxEnt							
	TP	TN	FP	FN	Pre	Rec	F1	Acc
Baseline (B)	79	6260	**9**	30	89.77	72.48	80.20	99.39
Syntactic & Temporal (S)	81	**6267**	2	28	**97.59**	74.31	84.38	99.53
B+Knowledge-base (K)	85	6266	3	24	96.59	77.98	86.29	99.58
B+Structural (St)	77	6260	**9**	**32**	89.53	70.64	78.97	99.36
S+K+St	**87**	6265	4	22	95.60	**79.82**	**87.00**	**99.59**

Table 4. Performance evaluation. TP: True positive, TN: True negative, FP: False positive, FN: False negative; Pre: Precision, Rec: Recall, F1: F-1 measure, Acc: Accuracy; B: Baseline features, S: Syntactic and temporal features, K: Knowledge-base features, St: Structural features. The highest value for each column is in boldface.

Table 5 includes the top ranked 25 features from different feature types and their combinations. From the performance values presented in Table 4 and the feature information presented in Table 5, we made the following observations on our feature types.

Syntactic and Temporal Features: When compared to baseline features, syntactic & temporal features increased precision by ~8%, recall by ~2%, and F-score by ~4%. Among the feature subtypes, the *bigrams* identified many interesting features such as "*is recommended", "follow up", "considered with"* which were listed among the top ranked features.

Rank	B	S	B+K	S+K+St
1	recommended (unigram)	if_IN (POS)	recommended (unigram)	TENSE_VB (tense)
2	if (unigram)	is recommended (bigram)	if (unigram)	is_recommended (bigram)
3	recommend (unigram)	recommended_VBN (POS)	C0750591 (UMLS Concept)	recommended_VBN (POS)
4	further (unigram)	VB (Tense)	recommend (unigram)	if_IN (POS)
5	be (unigram)	recommend (stemmed verb)	be (unigram)	recommend (stemmed verb)
6	correlation (unigram)	recommend_VB (POS)	correlation (unigram)	recommend_VB (POS)
7	considered (unigram)	evaluation_NN (POS)	C0439231 (UMLS Concept)	T060 (Semantic Type)
8	evaluation (unigram)	consid (stemmed verb)	months (unigram)	C0750591 (UMLS Concept)
9	months (unigram)	ultrasound_NN (POS)	T059 (Semantic Type)	consid (stemmed verb)
10	consider (unigram)	recommended . (bigram)	with (unigram)	recommended_. (bigram)
11	ultrasound (unigram)	further evaluation (bigram)	T060 (Semantic Type)	correlation_with (bigram)
12	biopsy (unigram)	further_JJ (POS)	further (unigram)	correlation_NN (POS)
13	mri (unigram)	follow-up_JJ (POS)	to (unigram)	further_evaluation (bigram)
14	follow-up (unigram)	correlation_NN (POS)	consider (unigram)	includesRADLEXConcept
15	with (unigram)	correlation with (bigram)	T058 (Semantic Type)	T058 (Semantic Type)
16	diagnosis (unigram)	be_VB (POS)	considered (unigram)	be_VB (POS)
17	suggest (unigram)	months_NNS (POS)	could (unigram)	T059 (Semantic Type)
18	advised (unigram)	mri_NNP (POS)	cta (unigram)	follow-up_JJ (POS)
19	could (unigram)	with ultrasound (bigram)	includesRADLEXConcept	C0439231 (UMLS Concept)
20	clinical (unigram)	with_IN (POS)	advised (unigram)	months_NNS (POS)
21	psa (unigram)	clinical_JJ (POS)	C1552861 (UMLS Concept)	with_IN (POS)
22	exclude (unigram)	consider_VB (POS)	helpful (unigram)	C1517331 (UMLS Concept)
23	helpful (unigram)	be considered (bigram)	ct (unigram)	consider_VB (POS)
24	sensitive (unigram)	considered_VBN (POS)	T028 (Semantic Type)	further_JJ (POS)
25	ct(unigram)	clinical_concern (bigram)	clinical (unigram)	evaluation_NN (POS)

Table 5. Top 25 ranked features for MaxEnt models with different feature combinations. B: Baseline features, S: Syntactic and temporal features, K: Knowledge-base features, St: Structural features. The type of each feature is given in parenthesis.

Knowledge-base Features: When added on the top of baseline features, the knowledge-base features increased the precision by ~7%, the recall by ~5, and F-score by ~6%. As can be seen from the B+K column of Table 5, all three subtypes had features among the top ranked 25 features (the binary feature *includesRADLEXConcept* was ranked 19th and there were four semantic types and three UMLS concepts). Especially, the semantic types *T059 - Laboratory Procedure* (ranked 9th), *T060 - Diagnostic Procedure* (ranked 11th), and *T068 - Human-caused Phenomenon or Process* (ranked 15th) successfully grouped concepts related to radiology tests and procedures.

Structural Features: Structural features did not change the precision and decreased the recall by ~2%. For positive recommendation class prediction, negative weights were assigned to all three section types; *Clinical Information*, *Description*, and *Impression*. The weight for *Impression* was closer to zero compared to the other two section types.

All Feature Types Combined: When all feature types combined, the precision increased by ~6%, the recall increased by ~7%, and F-score increased by ~7% compared to baseline performance values. As can be seen from the last column of Table 5, both syntactic & temporal (S) and knowledge-base (K) sub-types have features in the top ranked 25 features. This indicated both feature types were capturing information necessary for the classifier to make a decision.

Error Analysis

We analyzed the false positive and false negative sentences identified by the best performing classifier with the combined features and made the following observations.

False Positives: The best performing classifier resulted four false positives. One source of error is that our current text processing process does not include negation analysis. This resulted in the false identification of negated-recommendation sentences as positives. For instance, our classifier identified the following two sentences *"No ultrasound follow-up recommended"* and "*No further imaging evaluation or follow-up is necessary*" as recommendation sentences due to the highly weighted features such as *ultrasound, follow-up,* and *recommended*. Including negation analysis into the process will eliminate such errors and increase the overall precision.

Another source of error was due to our definition of critical recommendation, which requires that the recommendation is made by the radiologist of the current report as an advice to the referring clinician. Our classifier identified "*He was advised to go to the emergency room for further evaluation*" as a recommendation sentence. Although this sentence includes recommendation information, it was not labeled as positive in our gold standard corpus. We investigated this example in detail by checking the surrounding context in the actual radiology report. In this particular case, the radiologist identified multiple hypo-echoic lesions in the superior right lobe of the liver and noticed the patient was having fevers and elevated white blood cell count in his clinical records. The radiologist informed the patient that those lesions could represent hepatic abscesses and advised him to go to the emergency room. The patient refused and left the institution against medical advice. The sentence under question recorded a recommendation that had been given to the patient, not a request by the radiologist for the clinician to perform further investigation. Therefore, according to our definition of critical recommendation, this sentence was annotated correctly in the gold standard as a negative recommendation sentence. Our classifier mistakenly identified the sentence as positive due to strong features such as *further evaluation*. This example shows that identifying a critical recommendation is much harder than identifying a general recommendation that does not have that particular meaning.

False Negatives: The majority of the false negatives were boundary decisions where the negative and positive prediction probabilities were very close to each other. For example, for the false sentence "*Continued follow-up imaging at 6-month intervals is advised*", the positive class prediction probability was 0.48 and negative class prediction probability was 0.52. Although the sentence included strong features such as *follow-up* and *advised*, the evidence provided by the features was not enough to classify it as a positive sentence.

The main reason for the false negative cases was our limited training set. We had 109 positive recommendation sentences in our corpus and with such a small number of labeled examples the trained models could not capture the complete characteristics of the recommendation sentences.

Conclusion

In this paper, we described a text processing approach to identify the radiology report sentences that involve critical recommendations. We tested various feature types to increase the performance of our tools. The best F-score (87%) was achieved with the combination of syntactic & temporal, knowledge base and structural features.

There are various ways that our text processing approach could help clinical practice to improve patient safety. Firstly, if our critical recommendation sentence extraction approach is integrated into the radiology report generation system, it can be used to remind the radiologist to contact the referrer with a verbal recommendation. Such a verbal communication process does not uniformly occur at present. Secondly, if the extracted critical recommendation information is stored as part of radiology report index, this information can be later used to generate automatic e-mails that include the reports with critical recommendations to be sent to the referrers. Thirdly, visual cues (e.g., highlighting recommendation sentences and highlighting the report title in the list of imaging reports in the patient's EMR record) can be used to increase the visibility of recommendations in radiology reports. Visual cues such as this may reduce risk that recommendations are overlooked or forgotten.[22] Such a process would reduce the reliance solely on radiologist-to-referrer verbal communication and the chance of critical recommendations being overlooked by clinicians. Lastly, all reports containing critical recommendations can be entered on a follow-up recall system to ensure the essential follow-up investigations occurred, or were considered by the clinical team but deemed unnecessary at a later date, ensuring "closure of the loop." We plan to adopt our text processing approach to be used in these four ways, and to assess its impact on quality and safety of patient care. Moreover, reducing overutilization of medical imaging has been identified as a means of reducing growth in health care costs.[23] Enhancing our text processing approach to assess the variability in recommendation rates between radiologists may assist in reducing unnecessary radiologist recommendations for further costly investigations.

For future work, we will focus on the following three areas. First, the current system is trained on a small data set annotated by a single annotator. This was the main limitation of our study and in our future experiments, we plan to increase the training size with multiple annotators and improve classification accuracy. Second, the current system detects the section headers with manually crafted rules. Because our data was retrieved from the Harborview Medical Center Radiology Information System, the rule set we created captures only the characteristics of our institution's reports and requires modification if applied to other institutions' reports. We will explore various methods to make the section chunker more general. Third, in this paper, we focused on identifying critical recommendations; a remaining issue in radiology is a means of identifying various critical results in radiology reports. In the future, we plan to extend our work on detecting critical recommendations in radiology reports to identify critical test results in radiology reports.

Acknowledgements

This project was supported in part by an Institute of Translational Health Sciences Small Pilot Grant and the Patient Safety Innovations Program of the University of Washington.

References

1. Hussain S. Communicating Critical Results in Radiology. J Am Coll Radiol. 2010; 7(2):148-51.
2. American College of Radiology (ACR). ACR practice guideline for communication of diagnostic imaging findings. Accessed: July 1st, 2011. Available at: http://www.acr.org/SecondaryMainMenuCategories/quality_safety/guidelines/dx/comm_diag_rad.aspx
3. The Royal College of Radiologists. Standards for Communication of critical, urgent and unexpected significant radiological findings. Accessed: July1st, 2011. Available at: www.rcr.ac.uk/docs/radiology/pdf/Stand_urgent_reports.pdf
4. Towbin AJ, Hall S, Moskovitz J, Johnson ND, Donnelly LF. Creating a comprehensive customer service program to help convey critical and acute results of radiology studies. AJR Am J Roentgenol. 2011; 196(1):W48-51.
5. Lucey LL, Kushner DC. The ACR Guideline on Communication: To Be or Not to Be, That Is the Question. L Am Coll Radiol. 2010; 7(2): 109-114.
6. Kohn LT, Corrigan JM, Donaldson MS. To err is human: building a safer health system. Institute of Medicine, National Academy Press. 2000.
7. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A General Natural-language Text Processor for Clinical Radiology. JAMIA. 1994; 1: 161-174.
8. Friedman C, Johnson SB, Forman B, Starren J. Architectural requirements for a multipurpose natural language processor in the clinical environment. Proc. of Annu Symp Comput Appl Med Care. 1995; 347-351.
9. Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. Proc of AMIA Annu Fall Symp. 1996.
10. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp. 1997.
11. Hripcsak G, Austin JHM, Alderson PO, Friedman C. Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports. Radiology. 2002; (224):157-63.
12. Hersh W, Mailhot M, Arnott-Smith C, Lowe H. Selective Automated Indexing of Findings and Diagnoses in Radiology Reports. J Biomed Inform. 2001; 34: 262-73.
13. Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, Halpern EF, Thrall JH. Application of Recently Developed Computer Algorithm for Automatic Classification of Unstructured Radiology Reports: Validation Study. Radiology. 2005; 234:323-39.
14. Dang PA, Kalra MK, Blake MA, Schultz TJ, Halpern EF, Dreyer KJ. Extraction of Recommendation Features in Radiology with Natural Language Processing: Exploratory Study. AJR. 2008; 191:313-20.
15. Sistrom CL, Dreyer KJ, Dang PP, Weilburg JB, Boland GW, Rosenthal DI, Thrall JH. Recommendations for Additional Imaging in Radiology Reports: Multifactorial Analysis of 5.9 Million Examinations. Radiology. 2009; 253(2):453-61.
16. Chen H, Fuller SS, and Friedman C, Hersh W. Knowledge Management and Data Mining in Biomedicine. Springer. 2005.
17. Lewis DD. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. Proc. of ACM SIGIR 1992. pp. 37-50. 1992.
18. Toutanova K, Klein D, Manning CD, Singer Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. Proc. of HLT-NAACL 2003. pp 252-259, 2003.
19. Porter MF. An algorithm for suffix stripping. Program. 1980; 14(3):130-137.
20. Curtis P. Langlotz. RadLex: A New Method for Indexing Online Educational Materials. RadioGraphics. 2006; 26(6): 1595-97.
21. Berger AL, Pietra SAD, Pietra VJD. A maximum entropy approach to natural language processing. Journal of Computational Linguistics. 1996; 22(1):39-71.
22. Schiff GD. Medical Error: a 60-year-old man with delayed care for a renal mass. JAMA. 2011; 305(18):1890-8.
23. Hendee WR, Becker GJ, Borgstede JP, Bosma J, Casarella WJ, Erickson BA, Maynard CD, Thrall JH, Wallner PE. Addressing overutilization in medical imaging. Radiology. 2010; Oct;257(1):240-5.