Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A text processing pipeline to extract recommendations from radiology reports

Meliha Yetisgen-Yildiz [a,c,*], Martin L. Gunn [b], Fei Xia [c,a], Thomas H. Payne [d]

[a] Biomedical & Health Informatics, School of Medicine, University of Washington, Seattle, WA, United States
[b] Department of Radiology, School of Medicine, University of Washington, Seattle, WA, United States
[c] Department of Linguistics, University of Washington, Seattle, WA, United States
[d] Information Technology Services, School of Medicine, University of Washington, Seattle, WA, United States

ARTICLE INFO

Article history:
Received 7 May 2012
Accepted 30 December 2012
Available online 24 January 2013

Keywords:
Natural language processing
Recommendation identification
Section segmentation

ABSTRACT

Communication of follow-up recommendations when abnormalities are identified on imaging studies is prone to error. The absence of an automated system to identify and track radiology recommendations is an important barrier to ensuring timely follow-up of patients especially with non-acute incidental findings on imaging examinations. In this paper, we present a text processing pipeline to automatically identify clinically important recommendation sentences in radiology reports. Our extraction pipeline is based on natural language processing (NLP) and supervised text classification methods. To develop and test the pipeline, we created a corpus of 800 radiology reports double annotated for recommendation sentences by a radiologist and an internist. We ran several experiments to measure the impact of different feature types and the data imbalance between positive and negative recommendation sentences. Our fully statistical approach achieved the best *f*-score 0.758 in identifying the critical recommendation sentences in radiology reports.

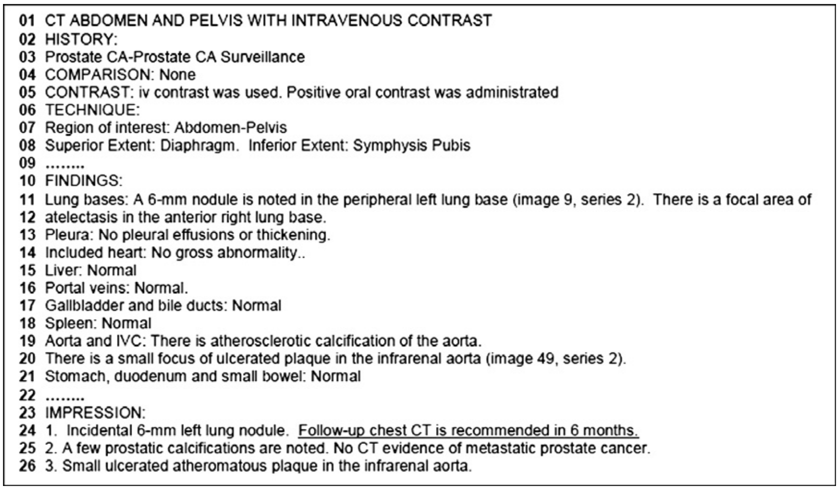
1. Introduction

Use of imaging technologies within healthcare delivery organizations has grown dramatically over the last decade, providing previously unavailable diagnostic and screening capabilities [1]. At the same time, growth in the number of reports and images generated contribute to growing challenge to optimally use clinical information while not being overwhelmed by it. In addition to providing reports addressing questions posed by ordering providers, radiologists often identify unexpected incidental findings that may pose a significant health risk to the patient in the short or medium term [2]. Professional societies have assisted in this effort by defining what types of findings justify designation as critical results, specified how these results should be communicated, and provided guidelines for the follow-up of incidental finding communication [2–4]. However, there is an unmet need to provide an automated system infrastructure to track and identify both critical test results and unexpected findings as these unexpected findings have been shown to “fall through the cracks” alarmingly frequently [5–7]. For example, the imaging report presented in Fig. 1 contains an incidental finding of a suspicious pulmonary nodule in a patient whose study was ordered because of another reason, which is accompanied by an ‘incidental recommendation’ that the study be repeated within 6 months because of concern for a malignant neoplasm (Fig. 1 – Line 24). These incidental findings may occasionally be communicated to the ordering clinician by phone or in person by the radiologist, but are usually written in the body of the radiology report [8]. However, because of patient mobility and frequent provider hand-offs, 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 may not be the same as the person who ordered the imaging study [5]. A study examining electronic transmission of critical radiology results that did not require recipients to acknowledge receipt of the notification, found that physicians failed to acknowledge the result in more than one third of cases, and in 4% of cases, the result was lost to follow-up 4 weeks after the test [6]. 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 electronic medical record (EMR). A recommendation will not be considered by the treating clinician if it is not seen.

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 [9]. The Joint Commission reported that up to 70% of sentinel medical errors were caused by communication errors [10]. Communication problems are at least

[*] Corresponding author. Address: Department of Biomedical Informatics and Medical Education, University of Washington, Box 357240, Seattle, WA 98195, United States. Fax: +1 206 221 2671.
E-mail address: melihay@uw.edu (M. Yetisgen-Yildiz).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.12.005



CT ABDOMEN AND PELVIS WITH INTRAVENOUS CONTRAST
HISTORY:
Prostate CA-Prostate CA Surveillance
COMPARISON: None
CONTRAST: iv contrast was used. Positive oral contrast was administrated
TECHNIQUE:
Region of interest: Abdomen-Pelvis
Superior Extent: Diaphragm. Inferior Extent: Symphysis Pubis
....
FINDINGS:
Lung bases: A 6-mm nodule is noted in the peripheral left lung base (image 9, series 2). There is a focal area of
atelectasis in the anterior right lung base.
Pleura: No pleural effusions or thickening.
Included heart: No gross abnormality..
Liver: Normal
Portal veins: Normal.
Gallbladder and bile ducts: Normal
Spleen: Normal
Aorta and IVC: There is atherosclerotic calcification of the aorta.
There is a small focus of ulcerated plaque in the infrarenal aorta (image 49, series 2).
Stomach, duodenum and small bowel: Normal
....
IMPRESSION:
1. Incidental 6-mm left lung nodule. Follow-up chest CT is recommended in 6 months.
2. A few prostatic calcifications are noted. No CT evidence of metastatic prostate cancer.
3. Small ulcerated atheromatous plaque in the infrarenal aorta.

Fig. 1. Example radiology report.

a causative factor in up to 80% of medical malpractice cases, and for radiologists, miscommunication is the second most common cause for malpractice lawsuits, after "failure to diagnose" [11].

The goal of our research is to identify clinically important follow-up recommendations so that the reports can be flagged visually and electronically so that 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 based on natural language processing (NLP) and machine learning to identify the sentences that involve clinically important recommendation information in radiology reports.

In this research study, we defined *clinically important 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 *clinically important recommendation* unless specified otherwise.

2. Related work

In the clinical NLP domain, radiology reports have been widely studied by various researchers [12–17]. 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 [12,13]. Jain et al. used MedLEE to encode the clinical information in chest and mammogram reports to identify suspected tuberculosis [14] and breast cancer [15] 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 [17]. In this paper, our overall goal is to identify sentences that include clinically important recommendation information in radiology reports.

The problem of identification of recommendation information in radiology reports has also been previously studied by other researchers [18–20]. 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 [18]. In that study the researchers did not analyze the documents at the sentence level. The same research group performed a similar analysis on a database of radiology reports covering the years 1995–2004 [19]. 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 were then 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 0.945 precision in identifying temporal phrases, and 0.932 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 [20]. 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 a previous study, we built a statistical recommendation sentence identification approach [21]. The main focus of that study was to define features that capture the characteristics of recommendation sentences. We ran several experiments with different combinations of the defined features to improve the overall classification performance. The pipeline presented in this paper is a continuum of that research. There are many contributions of this paper. First, we present a novel statistical section segmentation approach that chunks a given radiology report into its main sections. Second, we create a new gold standard that is double annotated by a radiologist and an internal medicine physician, which allows us to analyze disagreement between annotators. Third, we explore the effects of high dimensional feature space and the effects of data imbalance between positive and negative classes on the classification performance.

3. Methods

The main components of our recommendation extraction pipeline are depicted in Fig. 2. The pipeline includes (1) a section segmenter that divides a given radiology report into its main sections, (2) a sentence segmenter that identifies the sentence boundaries in the identified sections, and (3) a binary classifier that

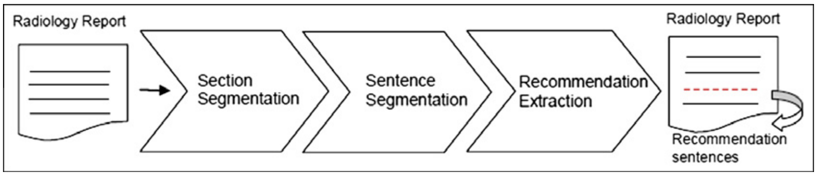


Fig. 2. Recommendation extraction pipeline.

labels each identified sentence as a positive or negative recommendation sentence. In the following sections, we will explain each component in detail.

3.1. Dataset

In order to develop and evaluate our system, we created a corpus of radiology reports composed of 800 de-identified radiology reports extracted from the Harborview Medical Center radiology information system. The reports had been generated by 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 radiography, computer tomography (CT), ultrasound and magnetic resonance imaging (MRI). All the reports in our dataset were created as structured template reports and edited later by the radiologists for the details and finding of the imaging tests. The radiologists used voice recognition systems to edit the reports. 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.

Two annotators, one radiologist (M.L.G., second author of the paper) and one internal medicine physician (T.H.P., fourth author of the paper), reviewed each of the 800 reports. The distribution of the reviewed reports across imaging modalities is listed in Table 1.

Table 1
Distribution of radiology reports.

Imaging modality	Frequency
Computer tomography (CT)	486
Radiograph	259
Magnetic resonance imaging (MRI)	45
Ultrasound	10

The annotators marked the sentences that include recommendations in the following two rounds.[1] Having a radiologist and an internal medicine physician enabled us to capture the views of the producers and the consumers of the radiology reports on recommendations sentences.

In the first round, there were no annotation guidelines other than the definition of recommendation. Each annotator annotated the data independently from each other. In the second round, each annotator went over the instances (an instance is a sentence in corpus) that received different labels in the first round and wrote a note to explain the rationale for his labeling; then he read the rationale written by the other annotator and relabeled the sentences if he agreed with the other annotator's rationale. As an example case, one annotator labeled the sentence "*If clinically indicated, pelvic ultrasound could be performed in 4 to 6 weeks to document resolution*" as recommendation, but the other annotator did not because he thought the author was hedging. In the second round, both annotators agreed that the sentence did not include recommendation.

With the two rounds of double annotations, we calculated inter-annotator agreement (IAA) for each round. The results are shown in Table 2. The IAA was pretty low for the first round (0.757). Going through the second round improved the IAA significantly (0.974). We used the 113 sentences agreed by the two annotators in the second round as the gold standard for the recommendation sentence identification task.

Table 2
IAA for the Radiology Corpus (C1). The corpus has 800 documents and 18,747 sentences in total. The "A1" and "A2" columns show the numbers of recommendation sentences (i.e., positive sentences) marked by the two annotators; the "Agreed" column shows the number of positive sentences marked by both annotators; P/R/F scores are precision, recall, and *f*-score for identifying positive sentences when A2's annotation is treated as gold standard and A1's annotation is treated as system output; "kappa" is the kappa coefficient.

Round	A1	A2	Agreed	P/R/F	Kappa
1st	110	109	83	0.755/0.761/0.758	0.757
2nd	114	118	113	0.991/0.958/0.974	0.974

3.2. Section and sentence segmentation

Although radiology reports are in free-text, they are structured in terms of sections to describe clinical information (e.g. Fig. 1 – Line 2 "*HISTORY*:"), exam details (e.g. Fig. 1 – Line 4 "*COMPARISON*:", Line 5 – "*CONTRAST*:", Line 6 – "*TECHNIQUE*:"), findings (e.g., Fig. 1 – Line 10: "*FINDINGS*:"), and impression (e.g., Fig. 1 – Line 23 "*IMPRESSION*:"). Automatically segmenting and classifying clinical free-text into sections is an important first step to any information extraction task as it helps to ground the significance of the text within.

In prior research, we developed a rule-based approach to identify the section boundaries of radiology reports [21]. Because the rules we defined mainly relied on the heuristics that capture the characteristics of radiology reports of our institution, they were not generalizable. To overcome this problem, we applied machine learning approaches to identify the boundaries of the sections and their types [22]. We constructed an ontology of 11 section categories for radiology reports (Table 3). The ontology has been designed to cover typical radiology reports sections, as advised by a clinical expert. To build the ontology, we first created a list of sections by randomly selecting a small subset of 10 radiology reports from the corpus. With the help of the clinical expert, we (a) grouped similar sections together under general section categories, and (b) put rare or atypical sections under a catch-all category. We used this ontology to create an annotated corpus to train our section segmentation approach. We randomly selected 100 reports from our corpus and a single annotator annotated the reports for section boundaries and assigned each annotated section a category from

[1] M.L.G. and T.H.P. participated in manual annotation and error analysis of the study. They did not actively participate in the system design and development phases to prevent bias to the system performance.

Table 3
Section category ontology for radiology reports and category frequencies in the annotated set of 100 radiology reports.

Section header categories	Frequency	Percentage (%)
Clinical information		
Clinical history	99	16.7
Exam details		
Exam	8	1.4
Comparison	89	15
Contrast	22	3.7
Procedure	70	11.8
Findings		
Findings	100	16.8
Impression		
Impression	77	13.0
Attending statement	14	2.4
Other		
Document heather	104	17.5
Catchall	6	1.0
Combined	5	0.8
Total	594	100

Table 4
Features for line labeling in Step 1.

Type	Features
Text features	*isAllCaps, isTitleCaps, containsNumber, beginsWithNumber, numTokens, numPreBlanklines, numPostBlanklines, firstToken, secondToken, unigram*
Tag features	*prevTag, prevTwoTags, tagChainLength*

Table 5
Features for section labeling in Step 2.

Type	Features
Header features	*Same as Text features, only the header line is used*
Body features	*avgLineLength, numLines, docPosition, containsList, unigram*
Tag features	*prevTag, tagHistUnigram, tagChainLength*

the ontology. There were 594 sections annotated in the 100 radiology reports.

Our basic methodology for section segmentation is to classify each line in a document to indicate its membership to a section. Our classifier operates at the line level rather than the sentence-level, because the content of clinical records tends to be fragmentary and list based. Under this methodology, we built two separate models for section segmentation and classification. First, the section boundaries are identified by labeling each line with a B (beginning of section), I (inside of section), or O (outside of section) tag. We used the features listed in Table 4 to create the feature vectors for this step. Then, the unlabeled sections from the first step are passed to the second step, where a separate classifier is called upon to label each section with a section category. The features used for the second step are listed in Table 5.

We used Maximum Entropy (MaxEnt) [23] models for classification and used beam search to find a good tag sequence. We used the MALLET toolkit [24] with L-BFGS parameter estimation and Gaussian prior smoothing. The Gaussian prior variance was left at its default value (=1). To measure the overall performance, we ran 5-fold cross validation and achieved 0.93 precision, 0.91 recall, and 0.92 *f*-score, where a match means that the section identified by the system matches a section in the gold standard with respect to both section boundaries and section label.

Table 6
Frequencies of categories identified by the statistical section segmenter in our corpus of 800 radiology reports.

Section header categories	Frequency	Percentage (%)
Clinical information		
Clinical history	863	18.3
Exam details		
Exam	79	1.7
Comparison	763	16.0
Contrast	190	4.0
Procedure	526	11.0
Findings		
Findings	826	17.4
Impression		
Impression	550	11.6
Attending statement	106	2.2
Other		
Document heather	849	17.8
Catchall	6	0.1
Total	4758	100

Our section segmentater identified a total of 4758 sections in 800 radiology reports. The distribution of the sections is in Table 6. After identifying the report sections, we used the OpenNLP[2] sentence chunker to identify the boundaries of the sentences in the section bodies. We identified 18,747 sentences (113 labeled as positive recommendation sentence and 18,634 labeled as negative recommendation sentence) in the 4758 sections of 800 reports in our corpus.

3.3. Recommendation extraction

The main architecture of our recommendation extraction component is illustrated in Fig. 3. In the following subsections, we will explain the main steps of the process.

3.3.1. Feature extractor

To identify recommendation sentences, we first represented the content of sentences with a feature vector that captures the characteristics and content of the sentences. We used five types of features as listed in Table 7.

We used unigrams as the baseline representation. Under the category Ngram features, we represented the content of the sentences with bigrams and trigrams. For syntactic features, we identified the part-of-speech tags of the unigrams by using Stanford POS Tagger [25].

We defined knowledge-based features based on Unified Medical Language System (UMLS). In the latest version of UMLS [27], 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 [28], a tool created by NLM that maps the strings in free text to biomedical concepts in the UMLS. 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 features in our representation. For syntactic features, 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 sentences. We defined a binary feature to capture the presence of modal verbs based on the observation that the majority of the positive sentences included a modal verb. To group different forms of common verbs used in recommendation sentences such as *recommend*, *suggest*,

[2] OpenNLP. Available at: http://opennlp.apache.org/index.html.

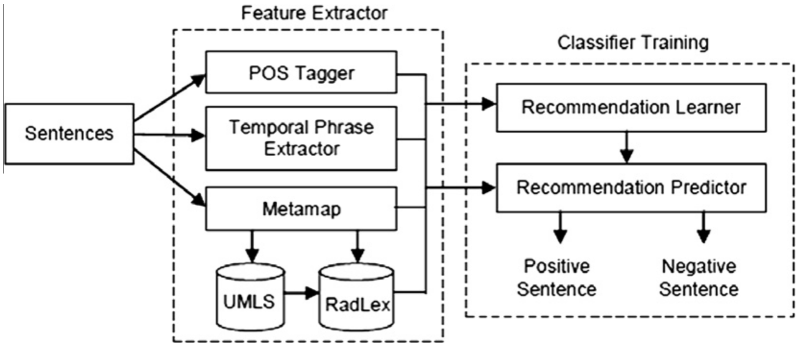


Fig. 3. Architecture for recommendation extractor.

Table 7
Features for sentence classification.

Feature type	Features
Baseline (B)	*unigram*
Ngram (N)	*bigram, trigram*
Syntactic (S)	*tense, stemmedVerb, includesModalVerb, includesTemporalPhrase*
Knowledge-based (K)	*umlsConcept*
Structural (St)	*SectionType*

consider, and *advise*, we stemmed the verbs with Porter stemmer [26] (e.g., Fig. 1 – Line 24: *recommended* ⇒ *recommend*).

Recommendation sentences usually appear under *Impression* sections (Fig. 1 – Lines 23–26). Such an observation indicated the potential importance of section header information in the classification decision. Under the group of structural features, we defined a feature based on section header of a given sentence to capture where the sentence is located within the report. For instance, the binary feature defined for the sentence "*Follow-up chest CT is recommended in 6 months*" (Fig. 1 – Line 24) is called *sectionHeader_Impression*.

3.3.2. Classifier training

For our recommendation sentence classification task, we used the implementation of MaxEnt algorithm available in MALLET toolkit [24].

4. Results

We evaluated the classification performance by using precision, recall, *f*-score, and accuracy performance metrics. Because there were a very limited number of positive sentences in our annotated corpus, we decided to use 5-fold cross validation to measure the performance of our classifiers.

4.1. Classification performance

We designed three groups of experiments to explore the effects of (1) the large feature space used to represent the dataset, (2) different feature types, and (3) data imbalance between positive and negative classes on the classification of recommendation sentences.

In our prior work, we have explored the effect of different feature types on the overall recommendation sentence classification performance and found out that the combined feature sets performed the best. Because we made revision to the gold standard based on double annotation and changed several components of our architecture, we decided to repeat the same experiment to see whether our prior findings were still valid. There were 109 positive sentences identified in the single annotated dataset used in prior work [21]. This number increased to 113 after double annotation as described in Section 3.1. However, only 91 of these 113 double annotated sentences were marked as positive in the previous gold standard created by a single annotator.

4.1.1. Feature set size experiments

As described in Section 3.3.1, we defined a rich set of features to represent the content of radiology report sentences. Table 8 summarizes the number of distinct features for each feature subtype used to represent the 113 positive sentences as well as the complete dataset. For some feature sub-types (e.g., *bigrams*, *trigrams*), the feature set sizes are quite high. Feature selection algorithms have been successfully applied in text classification in order to improve the classification accuracy. By significantly reducing the dimensionality of the feature space, they also improve the efficiency of the classifiers and provide a better understanding of the data [29,30].

In our experiments, before learning a model for recommendation sentences, we first built lists of ranked features from the training set for each fold. We applied feature selection only to feature types with a large number of features including *unigrams*, *bigrams*, *trigrams*, and *UMLS concepts*. We used statistical hypothesis testing to determine whether there is an association between a given feature and the two categories of our problem. Specifically, we computed the χ^2 statistics [31] which generate an ordering of features. Table 9 lists the top 10 unigrams, bigrams, trigrams, and UMLS concepts ranked by χ^2 statistics for one of the 5-folds. As can be observed from Table 9, many of the features are closely linked to the recommendation sentences.

Table 8
Feature set sizes.

Feature type	Feature sub-type	# of distinct features for 113 positive sentences	# of distinct features for the complete dataset
Baseline (B)	*Unigram*	402	4687
Ngram (N)	*Bigram*	856	30720
	Trigram	1005	51615
Knowledge-base (K)	*UMLS concept*	221	3987
Syntactic (S)	*Tense*	6	8
	stemmedVerb	55	564
	includesModalVerb	1	1
	includesTemporalPhrase	1	1
Structural (St)	*sectionType*	5	10

Table 9
Top 10 most informative unigrams, bigrams, trigrams and UMLS concepts for recommendation sentence classification according to χ^2 statistics when each feature type processed individually.

Rank	Unigram	Bigram	Trigram	UMLS concept
1	*If*	*further_evaluation*	*for_further_evaluation*	*C0034866-recommendation*
2	*further*	*for_further*	*further_evaluation_with*	*C1444656-indicated*
3	*evaluation*	*follow_up*	*if_clinically_indicated*	*C1719822-follow*
4	*recommended*	*is_recommended*	*helpful_for_further*	*C0589120-follow-up*
5	*follow*	*in_months*	*be_helpful_for*	*C0041618-ultrasound*
6	*recommend*	*evaluation_with*	*could_be_obtained*	*C0024485-MRI*
7	*consider*	*if_clinically*	*follow_up_ultrasound*	*C0332196-exclude*
8	*clinically*	*could_be*	*clinical_concern_for*	*C0205360-stability*
9	*ultrasound*	*if_there*	*if_there_is*	*C0220825-evaluation*
10	*indicated*	*be_obtained*	*ultrasound_could_be*	*C1301829-obtain*

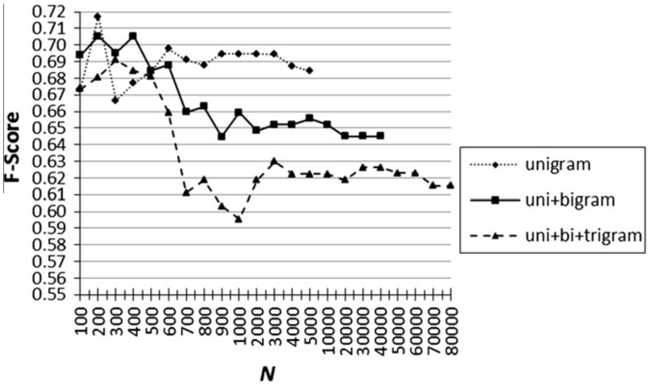


Fig. 4. Performance results for various subsets of top *N* word n-gram features ranked by χ^2 statistics.

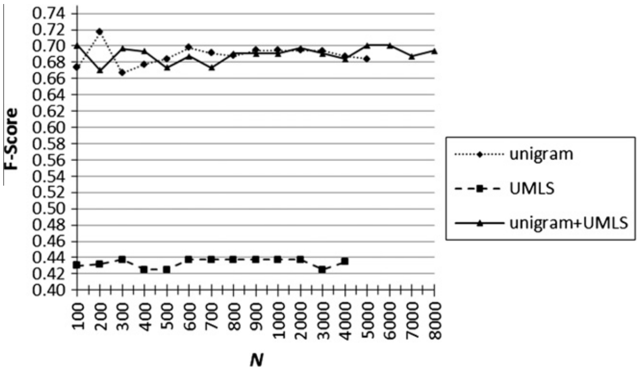


Fig. 5. Performance results for various subsets of top *N* unigram and UMLS concept features ranked by χ^2 statistics.

In the first set of experiments, we studied how the performance of our system evolves for various threshold values (*N*) on the different combinations of χ^2 ranked unigram, bigram, and trigram features. In order to rank the features associated with an experiment, we first pooled all features with the specific feature types used in the experiment under one set (i.e., unigram and bigram features were pooled for uni + bigram experiment). We constructed a contingency table for each feature from the pooled set and computed the χ^2 statistics. Once all features were ranked and their corresponding threshold value was established, we built a feature vector for each sentence. Specifically, given a subset of top *N* relevant features extracted from the ranked list of features, we considered in the representation of a given sentence's feature vector only the features from the subset of relevant features that were also found in the sentence. Therefore, the size of the feature space was equal to the size of the relevant features subset (*N*) whereas the length of each feature vector will be at most this value.

Fig. 4 shows the results of these experiments. The results are computed using the *f*-score which represents the harmonic mean of precision and recall. For unigrams, we considered 14 different values of *N* from a range of 100–5000. This number increased to 18 for uni + bigram experiment from a range of 100–35,000, and 22 for uni + bi + trigram experiment from a range of 100–80,000. As can be seen from the figure, in the overall, unigram performed better than uni + bigram and uni + bigram + trigram. The highest performance was achieved by unigram at *N* = 200 (precision = 0.835, recall = 0.628, *f*-score = 0.717).

In the second set of experiments, we compared unigrams with UMLS concepts for different values of *N*. As can be seen from Fig. 5, the performance of unigram is higher than UMLS concepts. When unigrams and UMLS concepts combined, the performance values are quite similar to those of only unigrams; however, the best performance was again achieved with unigram at *N* = 200. This finding can be explained by many of the strong unigram features (e.g., *if*, *further*) that are not in UMLS. These features capture the main characteristics of the recommendation sentences; therefore, adding UMLS concepts as features do not improve the overall classification performance. In the rest of the paper, we used unigrams at *N* = 200 as the baseline.

4.1.2. Feature type experiments

To understand the effect of each feature type, we added syntactic and structural features to the baseline features (unigrams at *N* = 200) individually and compared the classification performance. To see the effect of all features, we combined syntactic and structural features as the last feature combination, and compared its classification performance with the baseline performance.

Table 10 summarizes the results of the experiments. The accuracy was high for all the experiments since most instances were negative; therefore, precision, recall, and *f*-score were more informative for evaluation purposes. As can be seen from Table 10, each feature type increased the performance slightly when compared to the baseline. MaxEnt achieved the best precision with the structural features (B + St), and the best recall both with the syntactic (B + S) and combined features (B + S + St). The differences are minor and not statistically significant.

4.1.3. Data imbalance experiments

In our annotated dataset, the classes were imbalanced. There were only 113 (0.6%) sentences labeled as positive (minority class) and 18,634 (99.4%) sentences labeled as negative (majority class) for recommendation. The ratio of negative sentences over positive

Table 10
Performance evaluation. TP, true positive; TN, true negative; FP, false positive; FN, false negative; The highest precision, recall, *f*-score, and accuracy values are in bold.

Feature type	MaxEnt							
	TP	TN	FP	FN	Precision	Recall	*F*-score	Accuracy
Baseline – *unigram* (B)	71	18620	14	42	0.835	0.628	0.717	0.997
B + Syntactic (S)	73	18619	15	40	0.829	**0.646**	**0.726**	**0.997**
B + Structural (St)	72	18620	14	41	**0.837**	0.637	0.724	0.997
B + S + St	73	18618	16	40	0.820	**0.646**	0.723	0.997

sentences was 164.9 (18634/113 = 164.9). Because our corpus was randomly sampled from our radiology information system, this imbalance is *intrinsic*; in other words, the imbalance is a direct result of the nature of the data.

The fundamental problem with the imbalanced learning is that imbalanced data could lead to significant degradation of classification performance [32]. To understand the effect of data imbalance for our classification problem, we designed a series of experiments. Let P be the set of positive training instances and N be the set of negative training instances. For each k ($k = 1,\ldots,n$), we trained a classifier where the cardinality of N was equal to k times the cardinality of P ($|N| = k * |P|$).

Because we have a very small set of positive sentences, we use all positive sentences in each experiment; in contrast, the negative instances used for each k are randomly selected from the training set. We ran 165 experiments, one for each value of k between 1 and 165 (# of negative sentences/# of positive sentences = 18634/113 = 164.9). For each k, we ran a 5-fold cross validation to calculate the overall performance. We applied random negative sentence sampling to only training folds. The test fold included all the negative sentences so the experiments with different k are using exactly the same test set. We used all features types to create the feature vectors of the sentences.

Table 11 summarizes the results of the experiments where we evaluated the impact of data imbalance by using different ratios of negative and positive training sentences ($1 \leqslant k \leqslant 165$). Because we included all negative sentences in the test fold, accuracy is very high due to the imbalance between the classes. In contrast, precision, recall, and *f*-score are more informative. As can be seen from the table, when the sets are balanced ($k = 1$), recall is 0.9735 and precision is 0.2506. As k increases, recall decreases and precision increases gradually. At $k = 44$, precision and recall are closest to each other (precision = 0.7043, recall = 0.7168, *f*-score = 0.7105).

Fig. 6 includes the precision, recall, and *f*-score plots for different values of k. As can be seen from the graph, the performance change is sharper for smaller values of k and it stabilizes gradually as k increases.

4.2. Error analysis

We analyzed the false positive and false negative sentences identified by the classifier with highest *f*-score ($k = 9$) and made the following observations.

4.2.1. False positives

A main source of error was due to our definition of recommendation, which excludes hedge sentences that look very much like recommendations. To have one radiologist and one clinician as the two annotators enabled us to distinguish true recommendations from hedges cases in the gold standard, but our system does not have that expertise to make that distinctions perfectly. For instance, out of 51 false positives made by our best classifier ($k = 9$), 15 sentences (as listed in Table 12) are hedges, not true recommendations. Our classifier mistakenly identified those sentences as positive due to strong features such as *recommend*. These examples show identifying a clinically important recommendation is much harder than identifying a general recommendation.

Table 11
Performance evaluation with feature combination B($N = 200$) + S + St. k: class ratio, TP, true positive; TN, true negative; FP, false positive; FN, false negative; The highest precision, recall, and *f*-score values are in bold. The row with closest precision and recall is italicized ($k = 44$).

k	MaxEnt							
	TP	TN	FP	FN	Precision	Recall	*F*-score	Accuracy
1	110	18305	329	3	0.251	0.973	0.399	0.982
2	110	18433	201	3	0.354	0.973	0.519	0.989
3	105	18521	113	8	0.482	0.929	0.634	0.993
4	104	18536	98	9	0.515	0.920	0.660	0.994
5	103	18545	89	10	0.537	0.911	0.675	0.995
6	102	18565	69	11	0.537	0.903	0.718	0.996
7	101	18578	56	12	0.643	0.894	0.748	0.996
8	100	18581	53	13	0.654	0.886	0.752	0.996
9	100	18583	51	13	0.662	0.885	**0.758**	0.997
10	98	18584	50	15	0.662	0.877	0.751	0.996
20	92	18593	41	21	0.692	0.814	0.748	0.997
30	85	18594	40	28	0.680	0.752	0.714	0.996
40	82	18597	37	31	0.689	0.726	0.707	0.996
44	*81*	*18600*	*34*	*32*	*0.704*	*0.717*	*0.710*	*0.996*
50	80	18604	30	33	0.727	0.708	0.717	0.997
60	78	18606	28	35	0.736	0.690	0.712	0.997
70	77	18607	27	36	0.740	0.681	0.710	0.997
80	77	18608	26	36	0.748	0.681	0.713	0.997
90	77	18610	24	36	0.762	0.681	0.720	0.997
100	77	18611	23	36	0.770	0.681	0.723	0.997
165	73	18618	16	40	**0.820**	0.646	0.723	0.997

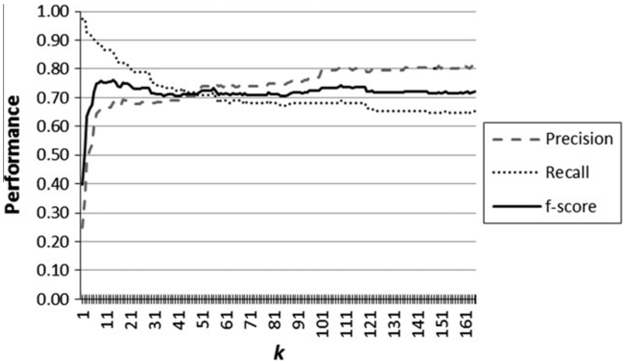


Fig. 6. Precision, recall, *f*-score curves. *k*: class ratio.

Another source of error is that our current text processing process does not include negation analysis. This results in the false identification of negated recommendation sentences as positives. For instance, our classifier identified "*No further follow-up is recommended*" as recommendation sentences due to highly weighted features such as *follow-up* and *recommend*.

4.2.2. False negatives

The main reason for the false negative cases was our limited training set. We had 113 positive recommendation sentences in

Table 12
False positive sentence examples.

1	No evidence of acute infarct, however, if there is further clinical concern, MRI is recommended
2	If further evaluation is clinically indicated, an ultrasound may be obtained
3	For more complete evaluation, additional imaging of the pelvis may be performed at incremental charge if desired
4	A CTA is recommended for further evaluation
5	This could be further evaluated with ultrasound
6	Nasogastric tube may be helpful for decompression if clinically indicated
7	If clinically indicated, nasogastric tube may be helpful for decompression
8	Recommend correlation with lab values
9	Recommend follow-up
10	Clinical correlation is recommended
11	Correlation with contents of aspirated fluid recommended
12	This could be further evaluated with MRCP or ERCP
13	We would generally follow 1 cm nodules for 2 years at 3–6 month intervals, however, if there are prior CTs available on this patient this would be helpful to avoid further scanning
14	If clinically indicated, pelvic ultrasound could be performed in 4–6 weeks to document resolution
15	If there is further clinical concern, a MR of the pelvis is a more sensitive study

our corpus and, with such a small number of labeled examples, the trained models could not capture the complete characteristics of the recommendation sentences. For example, for the false negative sentence "*Colonoscopy is recommended to rule out malignancy*", the positive class prediction probability was 0.49 and negative class prediction probability was 0.51. Although the sentence included strong features such as *recommend*, the evidence provided by the features was not strong enough to classify it as a positive sentence.

5. Conclusion

The main contribution of this paper is the text processing pipeline that identifies the radiology report sentences that involve clinically important recommendations. Our pipeline includes a novel statistical section segmentation approach that can be easily generalizable to radiology reports from other institutions. In addition, we showed that statistical feature selection improved the overall classification performance. In our experiments, we compared different feature set sizes for unigram, bigram, trigram, and UMLS concept features and found that top 200 unigram features ranked by χ^2 statistics achieved the best performance. Adding syntactic and structural features provided a modest gain, increasing *f*-score from 0.717 to 0.723 with B + S + St. Furthermore, we ran experiments to understand the effect of data imbalance between the positive and negative sentence classes. Our experiments showed that for lower values of *k*, we achieved very high recall but lower precision. As *k* increased, the recall gradually decreased and precision increased. We achieved the highest *f*-score of 0.758 at *k* = 9 (precision = 0.662, recall = 0.885).

For future work, we will focus on the following areas. First, the current system is trained on a small dataset double annotated by a radiologist and an internal medicine physician. The small dataset was the main limitation of our study; in our future experiments, we plan to increase the training size with multiple annotators and increase the classification performance. Second, because our primary research aim is to prevent medical errors due to missed recommendation information, recall is a more important measure than precision for this classification task. We achieved the higher levels of recall when the datasets were more balanced. In the future, we will investigate ways to increase precision for smaller values of *k* where recall is higher. Third, in this paper, we focused on identifying clinically important recommendations. For future work, we want to identify a particular type of recommendations within the set of clinically important recommendations – namely, the ones due to incidental findings – because they are the ones that are often missed by physicians because they often do not require immediate action and are not related to the medical concern for which the radiology test is requested by referring physicians.

Our ultimate goal is to create a production computing system that will be used in conjunction with our radiology information system and EMR to alert providers, highlight clinically important recommendations (especially the ones due to incidental findings) in the EMR, and develop a recall database to reduce the risk that important findings will be overlooked.

Acknowledgments

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] Hendee WR, Becker GJ, Borgstede JP, Bosma J, Casarella WJ, Erickson BA, et al. Addressing overutilization in medical imaging. Radiology 2010;257(1):240–5.
[2] Berland LL, Silverman SG, Gore RM, Mayo-Smith WW, Megibow AJ, Yee J, et al. Managing incidental findings on abdominal CT: white paper of the ACR incidental findings committee. J Am Coll Radiol 2010;7(10):754–73.
[3] American College of Radiology (ACR). ACR practice guideline for communication of diagnostic imaging findings. <http://www.acr.org/SecondaryMainMenuCategories/quality_safety/guidelines/dx/comm_diag_rad.aspx> [accessed 01.05.12].
[4] The Royal College of Radiologists. Standards for Communication of critical, urgent and unexpected significant radiological findings. <http://www.rcr.ac.uk/docs/radiology/pdf/Stand_urgent_reports.pdf> [accessed 01.05.12].
[5] Roy CL, Poon EG, Karson AS, Ladak-Merchant Z, Johnson RE, Maviglia SM, et al. Patient safety concerns arising from test results that return after hospital discharge. Ann Intern Med 2005;143(2):121–8.
[6] Singh H, Arora HS, Vij MS, Rao R, Khan MM, Petersen LA. Communication outcomes of critical imaging results in a computerized notification system. J Am Med Inform Assoc 2007;14(4):459–66.
[7] Schiff GD. Medical error: a 60-year-old man with delayed care for a renal mass. JAMA 2011;305(18):1890–8.
[8] MacMahon H, Austin JH, Gamsu G, Herold CJ, Jett JR, Naidich DP, et al. Guildelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. Radiology 2005;327(2):395–400.
[9] 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.
[10] Lucey LL, Kushner DC. The ACR guideline on communication: to be or not to be, that is the question. J Am Coll Radiol 2010;7(2):109–14.
[11] Berlin L. Failure of radiologic communication: an increasing cause of malpractice litigation and harm to patients. Appl Radiol 2012;39(1–2).
[12] Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. JAMIA 1994;1:161–74.
[13] Friedman C, Johnson SB, Forman B, Starren J. Architectural requirements for a multipurpose natural language processor in the clinical environment. In: Proc Annu Symp Comput Appl, Med Care; 1995. p. 347–51.
[14] Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. In: Proc AMIA Annu Fall Symp; 1996.
[15] Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In: Proc AMIA Annu Fall Symp; 1997.
[16] 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.
[17] 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.
[18] Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 2005;234:323–39.
[19] 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.
[20] Sistrom CL, Dreyer KJ, Dang PP, Weilburg JB, Boland GW, Rosenthal DI, et al. Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. Radiology 2009;253(2):453–61.
[21] Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. Automatic identification of critical follow-up recommendation sentences in radiology reports. AMIA Annu Symp Proc; 2011. p. 1593–602.

[22] Tepper M, Capurro D, Xia F, Vanderwende L, Yetisgen-Yildiz M. Statistical section segmentation in free-text clinical records. In: Proceeding of the international conference on language resources and evaluation (LREC), Istanbul, Turkey; May 2012.

[23] Berger AL, Pietra SAD, Pietra VJD. A maximum entropy approach to natural language processing. J Comput Linguist 1996;22(1):39–71.

[24] McCallum AK. A machine learning for language toolkit; 2002. <http://mallet.cs.umass.edu>.

[25] Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proc HLT-NAACL 2003; 2003. p. 252–9.

[26] Porter MF. An algorithm for suffix stripping. Program 1980;14(3):130–7.

[27] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucl Acids Res 2004;32(1):D267–70.

[28] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.

[29] Mladenic D, Grobelnik M. Feature selection on hierarchy of web documents. Decis Support Syst 2003;35:45–87.

[30] Wenqian W, Houkuan H, Haibin Z, et al. A novel feature selection algorithm for text categorization. Expert Syst Appl 2007;33:1–5.

[31] Manning CD, Schutze H. Foundations of statistical natural language processing. MIT Press; 1999.

[32] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowledge Data Eng 2009;21(9):1263–84.