Extraction and Analysis of Clinically Important Follow-up Recommendations in a Large Radiology Dataset

Wilson Lau[1], Thomas H. Payne, MD[2,3], Ozlem Uzuner, PhD[4], Meliha Yetisgen, PhD[1]
[1]Department of Biomedical and Health Informatics, [2]School of Medicine, [3]Information Technology Services, University of Washington, Seattle, WA
[4]Department of Information Sciences and Technology, George Mason University, Fairfax, VA

Abstract

Communication of follow-up recommendations when abnormalities are identified on imaging studies is prone to error. In this paper, we present a natural language processing approach based on deep learning to automatically identify clinically important recommendations in radiology reports. Our approach first identifies the recommendation sentences and then extracts reason, test, and time frame of the identified recommendations. To train our extraction models, we created a corpus of 1367 radiology reports annotated for recommendation information. Our extraction models achieved 0.93 f-score for recommendation sentence, 0.65 f-score for reason, 0.73 f-score for test, and 0.84 f-score for time frame. We applied the extraction models to a set of over 3.3 million radiology reports and analyzed the adherence of follow-up recommendations.

Introduction

With the dramatic rise in utilization of medical imaging in the past two decades, health providers are challenged by the optimal use of clinical information while not being overwhelmed by it. A radiology report is the principal means by which radiologists communicate the findings of an imaging test to the referring physician and sometimes the patient. Based on potentially important observations in the images, radiologists may recommend further imaging tests or a clinical follow-up in the narrative radiology report. These recommendations are made for several potential reasons: (1) radiologists may recommend further investigation to clarify the diagnosis or exclude potentially serious, but clinically expected disease; (2) radiologists may unexpectedly encounter signs of a potentially serious disease on the imaging study that they believe require further investigation; (3) radiologists may recommend surveillance of a disease to ensure an indolent course; or (4) they may provide advice to the referring physician about the most effective future tests specific to the patient's disease or risk factors. The reliance on human communication, documentation, and manual follow-up is a critical barrier to ensuring that appropriate imaging or clinical follow-up occurs. There are many potential points of failure when communicating and following up on important radiologic findings and recommendations: (1) Critical findings and follow-up recommendations not explicitly highlighted by radiologists: Although radiologists describe important incidental observations in reports, they may or may not phone an ordering physician. If these recommendations "fall through the cracks", patients may present months later with advanced disease (e.g., metastatic cancer). (2) Patient mobility: When patients move between services in healthcare facilities, there is increased risk during "hand-offs" of problems with follow-up of test result and continuity of care[1]. (3) Heavy workload of providers: Physicians and other providers have to deal with a deluge of test results. A survey of 262 physicians at 15 internal medicine practices found that physicians spend on average 74 minutes per clinical day managing test results, and 83% of physicians reported at least one delay in reviewing test results in the previous two months[2]. However, it is vital that these results, particularly if they are unexpected, are not lost to follow-up. In patients who have an unexpected finding on a chest radiograph, approximately 16% will eventually be diagnosed with a malignant neoplasm[3].

These examples indicate an opportunity to develop a systematic approach to augmenting existing channels of clinical information for preventing delays in diagnosis. The goals of our research are to: (1) build a gold standard corpus of radiology reports annotated with recommendation information, (2) build information extraction approaches based on deep learning to automatically identify recommendation information, and (3) apply the trained extractors to a large dataset of 3.3 million radiology reports created at University of Washington and Harborview Medical Center between 2008 and 2018 to analyze follow-up adherence rates.

In this research, we define a *follow-up 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. A follow-up recommendation can consist of multiple sentences. Figure 1 presents a radiology report with such a follow-up recommendation. In our annotation, we first labeled sentences with recommendation. For each identified recommendation, we also annotated the spans that describe (1) the reason for follow-up, (2) recommended test, and

(3) time frame. In figure 1, the recommendation is "*Given family history, would recommend repeat ultrasound in 4-5 weeks to evaluate fetal growth and complete anatomic survey. If unable to visualize fetal heart at that time, consider fetal echo*.", reason is "*to evaluate fetal growth and complete anatomic survey*", recommended test is "*ultrasound*", and time frame is "*4-5 weeks*".

IMPRESSION

Singleton pregnancy. Size consistent with dates. Anatomic survey limited by maternal body habitus and fetal position. Inadequate views of fetal heart and spine. *Given family history, would recommend repeat ultrasound in 4-5 weeks to evaluate fetal growth and complete anatomic survey*. If unable to visualize fetal heart at that time, consider fetal echo.

Figure 1. Example radiology report with recommendation information annotations.

Related Work

Automated information extraction using natural language processing (NLP) techniques has made patient information in clinical notes accessible for scientific research. Informatics for Integrating the Biology and the Bedside (i2b2) has been organizing NLP challenges on different types of clinical information extraction since 2006. These challenges included private health information de-identification[4], medical concept extraction[5], temporal information extraction[6], as well as medication information extraction[7]. Participants employed different NLP approaches including rule-based, machine learning and ensemble methods to address these tasks. Machine learning methods were usually based on statistical classification algorithms such as Support Vector Machines (SVMs), Maximum Entropy (MaxEnt) and Conditional Random Fields (CRFs)[8,9]. In recent years, neural networks have gained tremendous popularity, especially after several breakthroughs were accomplished by Hinton and Mikolov, T. et al.[10,11] and several deep learning libraries became publicly available. Clinical NLP researchers have also taken this opportunity and employed neural network modeling to deliver state-of-the-art performance. For instance, the best de-identification system in 2009 achieving 98% F-score was based on statistical learning with regular expression[12]. In 2016, Dernoncourt et al. were able to achieve similar performance using bidirectional LSTM and character embedding[13].

Prior research efforts on radiology follow-up recommendation detection are primarily based on rule-based and machine learning approaches. Dutta et al. employed lexicons matching heuristics to detect recommendations for incidental findings[14]. They compiled a set of lexicons which consisted of various inflectional morphemes of the same stem words. They went through three iterations of development and validation to fine tune their pattern matching algorithm. Chapman et al. adopted an algorithm, pyConTextNLP, to identify critical findings from radiology reports that were relevant to abdomen, chest, neuro and spine exams[15]. The algorithm used classification rules that were based on specific sentence structures in the reports. It also relied on a knowledge base that captured common biomedical terms in the target radiology imaging reports. Johnson et al. evaluated the ConText algorithm with a chest X-ray report and found that the algorithm misidentified two cases of negation and temporality in three sentences[16]. They proposed a heuristic approach to identify incidental findings based on regular expressions with patterns of lexicons. They argued that their approach could outperform processes that solely relied on radiologist annotations. However, their evaluation was based on a small and highly imbalanced dataset of 580 records (8.6% positive to negative ratio) and was only limited to X-ray, CT and ultrasound. Another lexicon based commercial system, LEXIMER (Lexicon Mediated Entropy Reduction), was used by Dang et al. to identify recommendations across different modalities. These authors analyzed the results using OLAP (Online analytical processing) technologies, a common approach in business intelligence and data warehousing[17,18]. LEXIMER parsed the reports into phrases and then weighted the phrases using hierarchical decision trees against a dictionary of lexicons[19]. Similarly, Mabotuwana et al extracted follow-up recommendations and associated anatomy from radiology reports with a keyword-based heuristic approach to identify recommendations in finding and impression sections of over 400 thousand radiology reports[20].The same group processed close to 3 million radiology notes to determine adherence rates to follow-up recommendations[21].

Domain adaptability is a major problem for rule-based and lexicon-based approaches as these methods require expert intervention to upkeep the logic of the rules and the dictionaries, which are often tailored to a specific problem and/or domain. Statistical NLP methods on the other hand do not require domain expert maintenance since the model automatically learns from annotated examples. Yetisgen-Yildiz et al. created a corpus of 800 radiology reports annotated with follow-up recommendations and developed a Maximum Entropy classifier for recommendation detection that achieved a best F-score of 87% based on a very rich set of features including ngrams, UMLS concepts, syntactic, temporal as well as structural features[22]. Similarly, Carrodeguas et al. created a corpus of 1000 randomly selected ultrasound, radiography, CT, and MRI reports annotated with follow-up recommendations. They trained

classifiers based on various supervised learning algorithms as well as recurrent neural networks. They achieved F-scores of 0.75 (random forest), 0.83 (logistic regression), and 0.85 (support vector machine).

Deep learning is not restricted by the lengthy process of handcrafted feature engineering usually required by traditional statistical NLP approaches for better performance. Instead, intricate distributed features are learned by adjusting model weights through backpropagation. Traditional methods suffer from word sense ambiguity and out-of-vocabulary tokens in clinical text which often contains misspellings, acronyms and foreign words. A common solution would be using dictionary of lexicons and gazetteers[24]. Deep learning can overcome these issues by the notion of transfer learning where the model is first trained on a larger dataset in a similar context and then fine-tuned on a smaller target dataset with limited number of annotated labels[25]. Another approach is to use character embeddings where the model is able to learn the morphological features of words, such as prefixes, suffixes, and any sub-token patterns to account for out-of-vocabulary words.

In this paper, we present a deep learning NLP system for extracting recommendation information from radiology imaging reports. We first develop a binary classifier based on Hierarchical Attention Networks[26] to identify follow-up recommendation sentences and then apply a state of the art deep neural named entity extraction system NeuroNER[27] to extract three entities: reason, test, time frame. These attributes help us understand why a follow-up recommendation is made by a radiologist to advise referring clinician for further evaluation. To our knowledge, this will be the first study that applied deep learning in the large scale to the problem of recommendation extraction in a dataset of 3.3 million radiology reports.

Methods

Datasets:

Pilot Corpus: In previous work, we created a corpus composed of 800 de-identified radiology reports extracted from the radiology information system of our institution. The reports represented a mixture of imaging modalities, including radiography computer tomography (CT), ultrasound, and magnetic resonance imaging (MRI). The distribution of the reports across imaging modalities is listed in Table 1.

Imaging modality	Number of reports
Computer tomography	486
Radiograph	259
Magnetic resonance imaging	45
Ultrasound	10
Total	800

Table 1. Distribution of reports in pilot corpus.

Annotation Guidelines: We annotated this dataset prior to defining different categories of follow-up recommendations. In this annotation task, we asked the annotators simply to highlight the boundaries of sentences that include any follow-up recommendations.

Annotation Process: Two annotators, one radiologist and one internal medicine specialist, went through each of the 800 reports independently and marked the sentences that contained follow-up recommendations. Out of 18,748 sentences in 800 reports, the radiologist annotated 118 sentences and the clinician annotated 114 sentences as recommendation. They agreed on 113 of the sentences annotated as recommendation. The inter-rater agreement measured in terms of F-score was 0.974.

Multi-institutional Radiology Corpus: We extended our pilot dataset of 800 reports with a much larger set of 3,301,748 radiology reports from two different institutions including the University of Washington Medical Center (1,903,772 reports) and Harborview Medical Center (1,397,976 reports) from year 2008 to 2018. University of Washington Human Subjects Division Institutional Review Board approved retrospective review of this dataset. Table 2 shows the distribution of radiology reports by modality in this larger dataset.

Annotation Process: We designed the annotation task to operate on two levels: sentence level and entity level. At the sentence level, the annotators mark the boundaries of recommendation sentences. At the entity level, the annotators mark three attributes of recommendation information presented in the marked sentences: (1) Test: the imaging test or clinical exam that is recommended for follow-up, e.g., *screening breast MRI or CT*, (2) Time frame: the recommended

time frame for the recommended follow-up test or exam, e.g., *1-3 weeks, 12 months*, and (3) Reason: the reason for the critical follow-up recommendation, e.g., *to assess the actual risk of Down's Syndrome*.

Imaging Modality	Number of reports
Angiography	53,658
Computed Tomography	706,908
Fluoroscopy	1,072
Magnetic Resonance Imaging	243,833
Mammogram	157,374
Nuclear Medicine	58,350
Portable Radiography	310,311
Positron emission tomography	1,799
Ultrasound	351,761
X-Ray	1,416,682
Total	3,301,748

Table 2. Distribution of reports in multi-institutional radiology corpus

Because manual annotation is a time-consuming and labor-intensive process, we could annotate only a small portion of our large radiology corpus. The percentage of reports that include recommendation sentences is quite low—about 15% at our institution. To increase the number of reports with recommendations in the annotated set, rather than randomly sampling, we built a high recall (0.90), low precision (0.35) classifier trained on the pilot dataset. The details of this baseline classifier can be found in our prior publication[28]. We ran our baseline classifier on un-annotated reports and only sampled from the ones identified as positive by our classifier for manual annotation. Because the baseline classification was high recall and low precision, the false positive reports could subsequently be corrected by our annotators. The filtering of reports using a classifier reduced the number of reports that our human annotators needed to review, thereby expediting the annotation process.

At the sentence level, one radiologist and one neurologist reviewed the classifier-selected reports with system generated follow-up recommendation sentences. The annotators corrected the system generated sentences and/or highlighted new sentences if needed.

At the entity level, one neurologist and one medical school student annotated the entities (reason for recommendation, recommended test, and time frame) in reports annotated in a previous stage at the sentence level with follow-up recommendations.

Inter-annotator Agreement Levels: At the sentence level, we measured the inter-annotator agreement on a set of 50 reports featuring at least one system-generated recommendation identified by our high recall classifier from a randomly selected collection of one thousand reports. Our annotation process required annotators to go over all sentences that were initially identified by the system as a recommendation. They could label the sentence as *Incorrect* if they believed the system had wrongly identified a recommendation sentence (false positive) or if they believed the system had missed the sentence (false negative). The inter-rater agreement levels were kappa 0.43 and 0.59 F1 score for the first iteration. To resolve the disagreements, we scheduled multiple meetings. One of our observations during those meetings was that none of the new recommendation sentences introduced by either annotator were identified by the other. In our review, both annotators agreed that the majority of the new recommendations introduced by the other were correct. We adjusted our annotation guidelines to add rules to help decide if and when a new sentence should be identified as a recommendation.

At the entity level, agreement levels were 0.78 F1 for reason, 0.88 F1 for test, and 0.84 F1 for time frame. Our final annotated corpus contained 597 positive examples of recommendation sentences and 11787 negative examples of recommendation sentences from 567 radiology reports. At the entity level there were 735 test, 173 time frame and 545 reason entities in the final corpus.

Approach:

Recommendation extraction: To identify sentences that include recommendation information, we first chunk reports into sentences with NLTK sentence tokenizer. Table 3 shows the distribution of sentences by image modality on the multi-institutional radiology dataset. As can be seen in the table, the length of reports varies across modalities.

Imaging Modality	Number of sentences	Average number of sentences per report
Angiography	1,504,939	28.05
Computed Tomography	18,109,590	25.62
Fluoroscopy	13,452	12.55
Magnetic Resonance Imaging	5,688,512	23.33
Mammogram	2,016,911	12.82
Nuclear Medicine	1,144,518	19.62
Portable Radiography	2,055,534	6.62
Positron emission tomography	41,423	23.03
Ultrasound	6,841,966	19.45
X-Ray	10,008,031	7.06

Table 3. Distribution of sentences by image modality in the multi-institutional radiology corpus

We defined our follow-up recommendation extraction task as a classification problem at the sentence level. We implemented our classifier based on Hierarchical Attention Networks (HANs)[26]. HAN is a neural model that employs a stacked recurrent neural network architecture. In particular, the weights of the hidden layers for each word are aggregated by an attention mechanism to form a sentence vector. The importance of each word in association with the outcome label is represented by the attention weight vector that can be learned by a layer of bidirectional Gated Recurrent Unit (GRU). The attention weight vector α is computed through a softmax function of the input context vector and a single hidden layer. Intuitively, the attention vector represents how important the word is in determining the outcome label. The sentence vector which is made up of these word attentions are then passed to another similar attention mechanism where the importance of sentences can also be learned by another layer of bidirectional GRU. The bidirectional nature of the encoders allows the contextual information in the input to be read in both directions and summarized. The hierarchical architecture allows the model to learn the context of a document by summarizing the context of its sentences, each of which in turn was summarized by its own words. The ability to selectively learn from local segments of texts to predict the outcome labels is a unique characteristic of attention mechanism in deep learning. This network model has been proven to be more effective than conventional statistical machine learning approaches in extracting clinical information[29]. In our annotated corpus, a single recommendation can comprise multiple sentences. We treated each recommendation as a document and trained the classifier to learn the relationship of the sentences within a recommendation. During inferencing, we classified each sentence separately and grouped consecutive positive predictions as one recommendation.

Hyperparameter optimization: We pretrained our word embeddings using Word2Vec on the entire 10 years of radiology dataset. We used grid search to find the best set of hyperparameters. Based on our preliminary experiments, we identified the range for each hyperparameter in the search space, which was also limited by available system memory: Word2Vec embedding dimension (100-300); number of bidirectional GRU unit on word encoder (100 - 500); number of bidirectional GRU unit on sentence encoder (100 - 500); drop out (0.3 - 0.5). We have also experimented with both Adam optimizer and SGD. Table 4 shows our best hyperparameter configuration.

Parameter	Value
word2vec embedding dimension	300
number of bidirectional GRU unit on word encoder	300
number of bidirectional GRU unit on sentence encoder	300
drop out	0.4
optimizer	Adam

Table 4. HAN hyperparameter configuration

We used 0.8/0.2(train/validation) split on the training dataset. We applied early stopping technique[30] to avoid overfitting with patience level set to 10 epochs. On each epoch, we evaluated the model based on the predicted F1 score on the validation set. The training would stop when no improvement was shown in the last 10 epochs.

Named Entity Recognition**:** We used Dernoncourt et al.'s NeuoNER[27] to process our annotated files in BRAT standoff format. The core of NeuroNER consists of two stacked layers of recurrent neural networks. The first layer is the *Character-enhanced token-embedding layer* in which the embedding of each word token is learned by a bidirectional LSTM from its character embedding. The resulting token embedding is then concatenated with our pretrained Word2Vec word embeddings to form an enhanced token embedding. These token embeddings are then processed by another layer of bidirectional LSTM, the *Label prediction layer*, to compute the probability vector of each word token being one of the entities. Finally, the sequence of probability vectors is sent to a feed-forward layer, the *Label sequence optimization layer*, to determine the predicted entity for each token by taking argmax of the probability vector, i.e., the entity label with the highest probability for each token. The character embedding captures the morphological features of word tokens and performs particularly well in handling morphemes, acronyms, misspellings and out-of-vocabulary tokens. It provides another level of word presentation that is not captured by sampling word co-occurrence as in Word2Vec and GloVe. This network architecture achieved state-of-the-art performance in identifying PHI information in i2b2 dataset and MIMIC dataset[13].

We used BIOES annotation (Begin, Inside, Outside, End, Single) to tag each token in the sequence and performed 5-fold cross validation on the training corpus. We pretrained our own word embeddings with Word2Vec on the multi-institutional radiology corpus of 3.3 million radiology reports.

Results

Recommendation extraction**:** We merged annotations from the pilot corpus and the multi-institutional radiology corpus to create one gold standard corpus that contains 693 positive sentences and 30429 negative sentences from a total of 1367 radiology reports. To understand the effect of data imbalance for our classification problem, we designed a series of experiments. Let P the set of positive training sentences and N be the set of negative sentences. For each k (k=1,…,n), we trained a classifier where the cardinality of N was equal to k times the cardinality of P. We performed 5-fold cross-validation at each value of K to obtain the average performance scores. We achieved the best 5-fold cross validation results at K=32 with 0.94 precision, 0.92 recall, and 0.93 f-score (true positive: 635, true negative: 11755, false positive: 39, false negative: 58). In previous work, for the same problem, we achieved 0.66 precision, 0.88 recall, 0.76 f-score with Max-Ent classifier with extensive feature engineering[28].

Named-entity recognition**:** Table 5 shows the token-based 5-fold cross validation results on the three entities.

Entity	Precision	Recall	F1
Reason	68.53	62.05	65.10
Test	74.20	71.48	72.71
Time frame	83.38	85.05	84.16

Table 5. Token level entity extraction 5-fold cross-validation results

Analysis of multi-institutional dataset: The recommendation extraction model predicted 685,912 recommendations in the total of 47,424,876 sentences. Table 6 shows the distribution of predicted recommendations and table 7 presents examples of extracted recommendation sentences by modality. 523,471 reports (15.9%) in the entire dataset included recommendations. As can be observed from Table 6, 98.02% of mammograms included a follow-up exam. For other modalities, percentages of reports with recommendations varied between 4.17% (portable radiography) and 25.66% (ultrasound). To evaluate the performance of our recommendation extraction model, we randomly selected 40 recommendations for top 5 modalities with highest recommendation percentages: mammograms (98.02%), ultrasound (25.66%), computed tomography (19.81%), positron emission tomography (18.68%), and Magnetic Resonance Imaging (14.32%) and manually validated their correctness. We identified 185 out of 200 of those recommendations as true positives which resulted a precision value (0.925) on the target dataset similar to our 5-fold cross validation result (0.94) on the annotated set.

We applied the NER model to extract entities from within the predicted recommendation sentences. Table 8 shows the distribution of predicted entities by modality. As can be observed from the example sentences presented in Table 7, not all recommendation sentences included reason, test, or time frame information. From 685,912 recommendations, the NER model extracted 250,840 (36.57%) reason, 528,040 (76.98%) test, and 216,128 (31.51%) time frame entities.

Imaging Modality	Number of recommendation sentences	Number of reports with recommendations (%)
Angiography	8455	7234 (13.48%)
Computed Tomography	193414	140066 (19.81%)
Fluoroscopy	103	100 (9.33%)
Magnetic Resonance Imaging	60954	34928 (14.32%)
Mammogram	210828	154255 (98.02%)
Nuclear Medicine	10141	7426 (12.73%)
Portable Radiography	13519	12951 (4.17%)
Positron emission tomography	472	336 (18.68%)
Ultrasound	109166	90266 (25.66%)
X-Ray	78860	75909 (5.36%)

Table 6. Number of predicted recommendations by modality

Imaging Modality	Example recommendation sentences
Angiography	*The patient will be followed up in the VIR clinic in approximately 2-3 weeks.*
Computed Tomography	*For a low risk patient, CT follow-up is recommended in 6 to 12 months. In the high risk patient, follow up is recommended at 3 to 6 months.*
Fluoroscopy	*Further evaluation with endoscopy is recommended.*
Magnetic Resonance Imaging	*BI-RADS category 6. Take appropriate action. MRI would be the best modality to assess response to neoadjuvant therapy.*
Mammogram	*Normal interval follow-up is recommended in 12 months.*
Nuclear Medicine	*Follow up nuclear medicine whole body scan is recommended in approximately 7 to 10 days after discharge.*
Portable Radiography	*A lateral radiograph or CT of the chest is recommended for further evaluation of this nodule.*
Positron emission tomography	*Follow up examination could be performed in 2 to 3 months to re-evaluate these lesions on PET.*
Ultrasound	*Recommend follow-up pelvic ultrasound in 2-3 months to evaluate for change.*
X-Ray	*Evaluation with weight bearing views is recommended.*

Table 7. Example recommendation sentences extracted from the dataset for each modality

Imaging Modality	Reason	Test	Time frame
Angiography	7,732	8,421	4,474
Computed Tomography	191,453	221,941	25,440
Fluoroscopy	159	125	7
Magnetic Resonance Imaging	41,136	68,452	20,679
Mammogram	24,998	250,605	162,421
Nuclear Medicine	11,895	12,476	974
Portable Radiography	15,292	15,725	367
Positron emission tomography	449	525	12
Ultrasound	82,371	134,233	36,827
X-Ray	73,383	65,115	2,894

Table 8. Number of predicted entities by modality

To understand the follow-up status of each identified recommendation, we performed a longitudinal analysis on the multi-institutional radiology dataset based on the information extracted by the NLP methods. To accomplish that, we first created a timeline of radiology reports for each patient based on the timestamps of reports in our dataset.

In our initial preliminary analysis, for each patient timeline we identified all reports with follow-up recommendations. We used the timestamps of the reports as the timestamp of the recommendations. For each identified recommendation, we checked whether a radiology test with the same modality occurred after the timestamp of the recommendation in the patient's timeline to roughly estimate the percentage of patients who stayed within the network of two hospitals in our dataset. Table 9 presents the results of this initial analysis.

Imaging Modality	Reports with follow-up recommendation	No following tests of same modality	Had following tests of same modality
Angiography	7234	2972 (41.08%)	4262 (58.92%)
Computed Tomography	140066	43698 (31.20%)	96368 (68.80%)
Fluoroscopy	100	84 (84.00%)	16 (16.00%)
Magnetic Resonance Imaging	34928	15791 (45.21%)	19137 (54.79%)
Mammogram	154255	45357 (29.40%)	108898 (70.60%)
Nuclear Medicine	7426	4131 (55.63%)	3295 (44.37%)
Portable Radiography	12951	3629 (28.02%)	9322 (71.98%)
Positron emission tomography	336	282 (83.93%)	54 (16.07%)
Ultrasound	90266	35067 (38.85%)	55199 (61.15%)
X-Ray	75909	22952 (30.24%)	52957 (69.76%)

Table 9. Number of patients who did / didn't have follow-up tests

Next, we used the entities extracted by our NLP methods. We first identified all reports that had recommendation with a time frame entity. The text segments of the time frame entities were then normalized to a common temporal expression using the Stanford temporal tagger (SUTime)[31]. SUTime normalizes the temporal phrases into a value (e.g., *3 months* = P3M, *1 year* = P1Y). Then using the timestamp of the recommendation and the normalized time frame value for follow-up, we projected the next radiologic test date for the patient. Because some projected dates are outside of the collected time range of the dataset, we considered those radiology encounters censored (18,338 records). If the recommended time consists of a range such as "*6 to 12 months*", we used the end of the range to project the next visit. Furthermore, a report could contain multiple follow-up recommendations (122,256 records). If the patient did not have any one of the follow-up encounters as recommended in the report, we considered no follow-up for that report. If the patient was late to any one of the recommended follow-up encounters in the report, we considered late follow-up for that report. Table 10 shows the number of patients who did not have a follow-up encounter as recommended by radiologist as well as those who had a follow-up earlier or later than the recommended time.

Imaging Modality	Reports with recommendation and projected time frame	No follow-up	Early follow-up	Late follow-up
Angiography	2075	759 (36.58%)	393 (18.94%)	923 (44.48%)
Computed Tomography	14506	5516 (38.03%)	4716 (32.51%)	4274 (29.46%)
Fluoroscopy	5	3 (60.00%)	0 (0%)	2 (40.00%)
Magnetic Resonance Imaging	8708	3393 (38.96%)	1736 (19.94%)	3579 (41.10%)
Mammogram	121716	27689 (22.75%)	19935 (16.38%)	74092 (60.87%)
Nuclear Medicine	349	143 (40.97%)	124 (35.53%)	82 (23.50%)
Portable Radiography	222	113 (50.90%)	62 (27.93%)	47 (21.17%)
Positron emission tomography	7	6 (85.71%)	0 (0%)	1 (14.29%)
Ultrasound	21083	8599 (40.79%)	5060 (24.00%)	7424 (35.21%)
X-Ray	976	354 (36.27%)	233 (23.87%)	389 (39.86%)

Table 10. Number of patients who had no follow-up / early follow-up / late follow-up

As can be observed from Table 10, mammograms had the highest follow-up rate (77%: 16% early, 61% late follow-up). This is expected as mammograms are commonly used as a screening tool to detect early breast cancer in women

and annual exam is recommended for women over 40yo. For the other modalities, the follow-up rates varied between 14% (positron emission tomography) and 64% (X-Ray).

Conclusion

The main contribution of this paper is the application of deep learning to identify clinically important recommendation information in radiology notes. We applied the trained models to multi-institutional dataset of 3.3 million radiology notes and presented our very preliminary analysis of recommendation follow-up adherence over a period of 10 years.

One of the limitations of our study was the size of the training set for recommendations. To achieve good performance, deep learning approaches require relatively larger dataset than traditional machine learning methods. Our labeled training corpus consists of only 1367 reports. The presented performance results are very promising. However, there is still room for improvement in recommendation extraction as well as NER tasks for reason, test, and time frame. We plan to annotate more reports to increase the size of our training set. We will also explore other deep learning methods including contextual embedding and transformer architecture such as BERT.

In our error analysis, we found that some of the time frame entities could not be normalized by Stanford's temporal tagger, such as "second trimester" in the recommendation "*Follow-up ultrasound is recommended in the early second trimester for further evaluation.*". 216,128 recommendations (32% of all recommendations) had time-frame entities and 169,647 (25% of all recommendations) of those with normalized time frames were included in the preliminary analysis. To utilize our entire dataset, we will build our own normalization algorithm for time frame entities. Additionally, we will automatically learn the recommended time frames for each modality from the data and use this knowledge to fill the missing time information for recommendations without time frame entities.

Our analysis did not utilize the extracted test and reason entities. We assumed the recommended test would be of the same modality of the original test with recommendation mentioned in its report. However in reality, recommended test may be of a different modality or of the same modality but with a different anatomy. In future work, we will extract the recommended anatomy in addition to other entity types. In addition, test, anatomy, and reason entities will be mapped to standardized dictionaries to enable a more comprehensive follow-up adherence analysis.

Acknowledgements

This publication was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR002319.

References

1. Callen J, Georgiou A, Li J, Westbrook JI. The safety implications of missed test results for hospitalised patients: a systematic review. BMJ Qual Saf. 2011;20(2):194-199.
2. Holden WE, Lewinsohn DM, Osborne ML, et al. Use of a Clinical Pathway To Manage Unsuspected Radiographic Findings. Chest. 2004;125(5):1753-1760.
3. Poon EG, Gandhi TK, Sequist TD, Murff HJ, Karson AS, Bates DW. "I Wish I Had Seen This Test Result Earlier!": Dissatisfaction With Test Result Management Systems in Primary Care. Arch Intern Med. 2004;164(20):2223-2228.
4. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform. 2015;58 Suppl:S20-29.
5. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc JAMIA. 2011;18(5):552-556.
6. Sun W, Rumshisky A, Uzuner O. Annotating temporal information in clinical narratives. J Biomed Inform. 2013;46 Suppl:S5-12.
7. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc JAMIA. 2010;17(5):514-518.
8. Patrick J, Li M. A Cascade Approach to Extracting Medication Events. In: Proceedings of the Australasian Language Technology Association Workshop 2009. Sydney, Australia; 2009:99–103.
9. Halgrim S, Xia F, Solti I, Cadag E, Uzuner Ö. Extracting Medication Information from Discharge Summaries. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents. Association for Computational Linguistics; 2010.
10. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504-507.

11. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. January 2013. ArXiv:1301.3781 [Cs].
12. Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. J Am Med Inform Assoc JAMIA. 2007;14(5):550-563.
13. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc. 2017;24(3):596-606.
14. Dutta S, Long WJ, Brown DFM, Reisner AT. Automated Detection Using Natural Language Processing of Radiologists Recommendations for Additional Imaging of Incidental Findings. Ann Emerg Med. 2013;62(2):162-169.
15. Chapman BE, Mowery DL, Narasimhan E, Patel N, Chapman W, Heilbrun M. Assessing the Feasibility of an Automated Suggestion System for Communicating Critical Findings from Chest Radiology Reports to Referring Physicians. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany: Association for Computational Linguistics; 2016:181–185.
16. Johnson E, Baughman WC, Ozsoyoglu G. Modeling Incidental Findings in Radiology Records. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. BCB'13. New York, NY, USA: ACM; 2013:940:940–940:945.
17. Dang PA, Kalra MK, Schultz TJ, Graham SA, Dreyer KJ. Informatics in Radiology. RadioGraphics. 2009;29(5):1233-1246.
18. Dang PA, Kalra MK, Blake MA, et al. Natural language processing using online analytic processing for assessing recommendations in radiology reports. J Am Coll Radiol JACR. 2008;5(3):197-204.
19. Dreyer KJ, Kalra MK, Maher MM, et al. Application of Recently Developed Computer Algorithm for Automatic Classification of Unstructured Radiology Reports: Validation Study. Radiology. 2005;234(2):323-329.
20. Mabotuwana T, Hall CS, Dalal S, Tieder J, Gunn ML. Extracting Follow-Up Recommendations and Associated Anatomy from Radiology Reports. Stud Health Technol Inform. 2017;245:1090-1094.
21. Mabotuwana T, Hombal V, Dalal S, Hall CS, Gunn M. Determining Adherence to Follow-up Imaging Recommendations. J Am Coll Radiol. 2018;15(3, Part A):422-428.
22. 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;2011:1593-1602.
23. Carrodeguas E, Lacson R, Swanson W, Khorasani R. Use of Machine Learning to Identify Follow-Up Recommendations in Radiology Reports. J Am Coll Radiol. 2019 Mar;16(3):336-343.
24. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994:235-239.
25. Lee JY, Dernoncourt F, Szolovits P. Transfer Learning for Named-Entity Recognition with Neural Networks. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). European Language Resource Association; 2018.
26. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical Attention Networks for Document Classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics; 2016:1480-1489.
27. Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics; 2017:97–102.
28. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform. 2013;46(2):354-362.
29. Gao S, Young MT, Qiu JX, et al. Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc JAMIA. November 2017.
30. Prechelt L. Early Stopping - But When? In: Orr GB, Müller K-R, eds. Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg; 1998:55-69.
31. Chang AX, Manning C. SUTime: A library for recognizing and normalizing time expressions. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). European Language Resources Association (ELRA); 2012.