# Clinically Significant Information Extraction from Radiology Reports

Nidhin Nandhakumar
Faculty of Computer Science,
Dalhousie University
Halifax, NS, Canada
nidhin.nandhakumar@dal.ca

Ehsan Sherkat
Faculty of Computer Science,
Dalhousie University
Halifax, NS, Canada
ehsansherkat@dal.ca

Evangelos E. Milios
Faculty of Computer Science,
Dalhousie University
Halifax, NS, Canada
eem@cs.dal.ca

Hong Gu
Department of Mathematics and
Statistics, Dalhousie University
Halifax, NS, Canada
hgu@dal.ca

Michael Butler
Department of Medicine,
Dalhousie University
Halifax, NS, Canada
mbbutler@dal.ca

## ABSTRACT

Radiology reports are one of the most important medical documents that a diagnostician looks into, especially in the emergency context. They provide the emergency physicians with critical information regarding the condition of the patient and help the physicians take immediate action on urgent conditions. However, the reports are in the form of unstructured text, which makes them time consuming for humans to interpret. We have developed a machine learning system to (a) efficiently extract the clinically significant parts and their level of importance in radiology reports, and (b) to classifies the overall report into *critical* or *non-critical* categories which help doctors to identify potential high priority reports. As a starting point, the system uses anonymized chest X-RAY reports of adults and provides three levels of importance for medical phrases. We used the Conditional Random Field (CRF) model to identify clinically significant phrases with an average f1-score of 0.75. The proposed system includes a web-based interface which highlights the medical phrases, and their level of importance to the emergency physician. The overall classification of the report is performed using the phrases extracted from the CRF model as features for the classifier. Average accuracy achieved is 85%.

## CCS CONCEPTS

**•Computing methodologies → Machine learning approaches;** *Classification and regression trees;* **•Applied computing → Health care information systems;**

## KEYWORDS

Information Extraction, Classification, Radiology Reports


*DocEng '17, September 4-7, 2017, Valletta, Malta.*

DOI: http://dx.doi.org/10.1145/3103010.3103023

## 1 INTRODUCTION

One of the key resources of information used by doctors (especially emergency physicians) are radiology reports [13] of the patients. The radiology report comprises key medical observations dictated by the radiologist when analyzing the patient's medical imaging reports (for example, x-rays) and these are automatically transcribed to text. It is the emergency physician who makes the decision on the treatment of the medical conditions. In the case of long radiology reports, the doctor may miss some of the key observations made by the radiologist. Another complexity in the processing of radiology reports is the presence of transcription errors.

Our proposed model tries to aid the emergency physicians in automatically identifying key medical observations from the radiology report, based on the criticality level of the medical phrases, using machine learning and a web-based visual interface. The system highlights the medical phrases on the fly, based on their criticality values for the doctors. We also classify the overall report to identify if the patient is in need of urgent treatment. We have listed our main contributions as the following:

- The design of a novel system which identifies the medical phrases and their associated criticality values and presents this information in a visual interface. In terms of performance, our proposed method is able to achieve similar performance to human annotators when identifying key phrases and their criticality level.
- The design of a Web-based tagging system which can be used by doctors for annotating the radiology reports to provide training data for the machine learning model.
- We have also managed to improve the accuracy of word segmentation and spelling correction algorithms and have tuned them for use in radiology reports.
- Designing a novel binary classification system for extracting radiology reports of critical- condition patients. The proposed approach was able to achieve better performance as compared to using 'bag of words' having tf-idf weights. We have managed to use a novel list of features for better classification of the radiology reports.

We start this paper with an overview of related models, or systems, which use radiology or similar medical reports for extracting

information from unstructured data. In Section 3, the overall view of our proposed model is introduced. The implementation details of each of the modules are mentioned in various subsections. We then compare and evaluate the performance of our system in Section 4. Finally, we analyze the type and cause of errors in the model and conclude the paper in Sections 5 and 6.

## 2 RELATED WORKS

Recently, Computational analysis of radiology reports gain a lot of attention. Most of the works focus on the identifying of specific medical conditions present in a given report and they usually deal with the classification task. Another generic area of research is the information extraction models which try to extract specific information such as medical recommendations or drug dosage from the radiology reports.

A system for identifying named entities from the radiology reports was the work done by Hassanpour [14]. The objective of this model was to identify the specific named entities from the radiology reports based on their information extraction model. They used a CRF-based model with several auxiliary features including POS tags and Radlex Lexicons [17] to identify entities of five different classes of *Anatomy*, *Anatomy Modifier*, *Observation*, *Observation Modifier*, and *Uncertainty*. They used 150 chest CT reports for training the system and reported an average f1-score [30] of 0.85.

Another system, Textractor [20] used the regular expressions to identify the medications and the reason for the prescription from patients' EHR (Electronic Health Record) files. The system uses the UMLS [5] concepts to identify the medication information and also uses the structure of medical reports to extract the reason for the prescriptions. Patrick and Li [24] used discharge files to identify the medication information such as Dosage, Mode, Frequency, Duration, Reason, and Context by using a hybrid machine learning and rule-based model. They used a combination of SVM [15] and CRF [31] models for identifying the entities and the rules used for final predictions.

In Information theory entropy reduction program [10] Dreyer used *Decision Trees* to extract the clinical findings and recommendations in the radiology report. However, the exact implementation details of their model is not provided by the author for replicating the results. Another work by Yetisgen [32] is a text processing pipeline for extracting recommendations from the radiology reports based on MEMM [19] model. However, the data set they used is highly unbalanced with 99% reports being negative. Another interesting model is the CTakes system [27] which identifies clinically significant phrases by the use of a combination of machine learning and rule-based models. However, the system's performance depends on the availability of up-to-date dictionaries and its performance is lowered as complexity increases.

Some of the more recent work includes the extraction of tumor information from radiology reports [33]. In this model, the authors were trying to extract tumor information for Hepatocellular carcinoma (HCC) disease. They used the CRF and MEMM models for extracting tumor's information such as tumor size, tumor count, and anatomical parts. They used a window size of 2 with a unigram model and limited the scope of the model to only 'findings' and 'impression' parts of the radiology report. The authors reported an f1-score of 0.74 in identification of tumor's information.

Our model tries to design a system that can extract clinical information without focus on any specific disease or clinical data. The information extracted can be used as key information for bigger models such as high level patient profiling systems and advanced machine learning tasks which use radiology data. We have selected a novel list of robust features in our proposed system.

## 3 THE PROPOSED METHOD

We implemented our system on a real-world anonymised radiology data set. The data included several spelling errors created by the automatic voice-to-text transcription, which had to be corrected before the data could be used for our machine learning model.

The overall structure of our model is shown in Figure 1. The main parts of our model include:

(1) Document Preparation
(2) Feature extraction
(3) Information Extraction
(4) Document Classification and
(5) Interface for Active Adaptive Learning

Each parts of the system is explained in detail in the following.

### 3.1 Document Preparation

Real world radiology reports are usually created by a voice-to-text processing system that creates text files based on the radiologist's dictation. In our model, for processing the reports effectively, we have to pre-process them before extracting the required information. The document-preparation module consists of two parts, a) the word segmentation module. b) the spelling correction module.

*3.1.1 Word Segmentation.* One of the most common errors in the radiology dataset is consecutive words that joined together. We implemented a word segmentation module to correct these joined-word errors. We used a probabilistic model [23] based on the Google trillion corpus [12]. This algorithm uses both unigrams and bigrams to generate the probabilistic values for each of the segmentations of the given word. The combination with a higher probabilistic value is selected as the corrected word. For example, If the word is 'isan' then, the model identifies that the words 'is' and 'an' separately produce better probability than 'isan' as a single word.

However, we cannot apply this algorithm 'out of the box', since we are dealing with medical terms. The occurrence of medical terms in the real world is much lower than common words. So the system would produce inaccurate results for most of the medical terms. For example, 'nabothian' would be segmented into 'na' + 'both' + 'ian', since these separate words are more common than 'nabothian'. For this reason we have modified the algorithm in two ways:

- We created a dictionary of unigrams and bigrams from the radiology dataset. Based on Radlex [17] and UMLS [5] ontologies, we extracted the medical and radiological terms from this dictionary. Because the occurrence of these terms are not sufficient for our task, we manipulated the word count for these terms based on their counts in the original Google n-gram corpus. We attached these terms with first

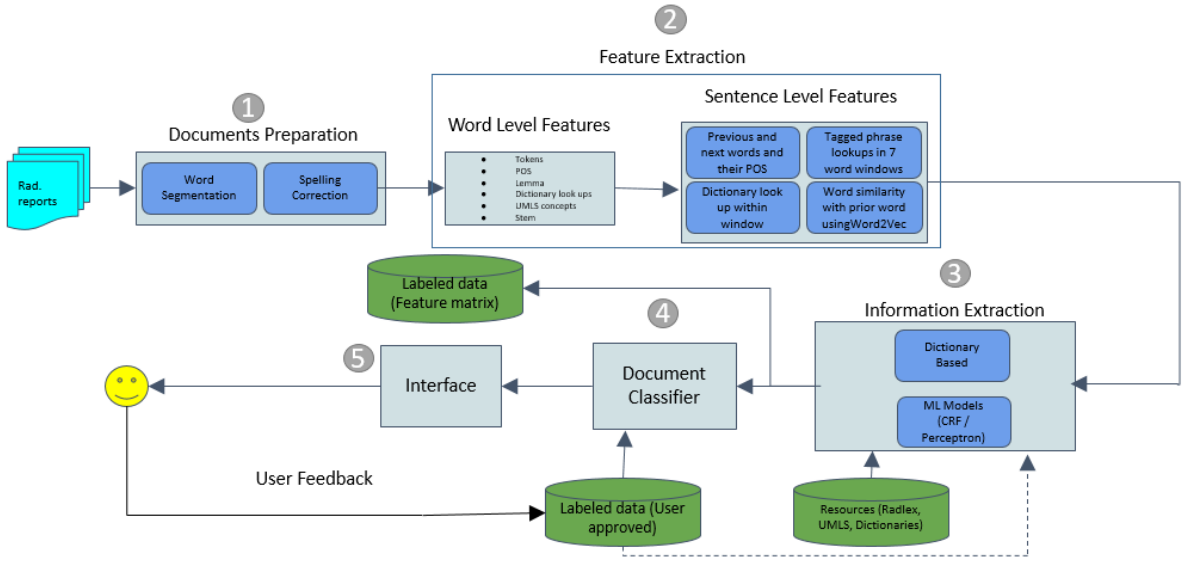


**Figure 1: The overall view of the proposed system. At first all reports will be preprocessed (1) then several word and sentence level features will be extracted (2). The Information Extraction module (3) used the extracted features for identifying important phrases with their level of importance. The Document Classifier (4) classifies reports into two critical and non-critical categories based on the information exacted from the previous step. The visual interface (5) provides the user the extracted information and then tries to incorporate the user feedbacks in the system.**

333,000 and 250,000 frequent unigrams and bigrams of Google trillion corpus.

- We also used Radlex [17] and UMLS [5] dictionaries which we created from the Radlex ontology and the UMLS Ontology. Each word was checked, in the UMLS and Radlex ontologies, to see if it was a valid medical term. Only terms which were not present in these dictionaries were processed for joined word error correction. This increased both the speed and accuracy of the word segmentation system.

*3.1.2 Spell Correction.* The spelling correction module was used for correcting normal spell errors occurred in the system. By analyzing the radiology reports, we found that the spelling errors were comparatively less frequent than the word segmentation errors. The spelling correction system is based on a on a probabilistic model [23].

However, since we are dealing with radiology data where spelling-error occurrence is less frequent, we created the word counts from the word-segmented radiology reports from the previous step. A word will not be checked for the spelling error if it present in the Radlex or UMLS ontologies.

## 3.2 Feature Extraction

This section explains in detail the auxiliary features used in training the machine learning models for extracting the clinically significant medical phrases. The first set of features are the word-level features discussed in detail in Section 3.2.1 and the second ones are the sentence-level features discussed in Section 3.2.2.

*3.2.1 Word-level Features.* Word-level features are auxiliary features extracted from word level syntactic and semantic analyses of reports. These features are used by the CRF model [16] for the information extraction and we explicitly defined them for enhancing the performance of the model. The various word-level features extracted are:

- Stem and lemma of the word: The *stem* is the core part of a word. For example, the stem of playing is play. The *lemma* is the canonical or dictionary form of the word.
- Part of speech: We used the MedPost/SKR part-of-speech tagger [29] to extract the POS tags for our words.
- Word length: length of the word (number of characters).
- Anatomy: This is a boolean flag value which is set if the given word is an anatomical word. The anatomy dictionary for this flag is generated from the Radlex [17] ontology.
- Suffix and prefix: We extract the first and the last two letters of a word as a two-letter prefix and suffix. We also use the first and the last three letters of the word as three-letter prefixes and suffixes, respectively.
- Critical level flags: This is a boolean flag value which is set if the given word is a high-critical, critical or non-critical word. This dictionary is created based on the tagged data set generated by the human tagger.
- Meta Label and Meta concept: This is the Meta Label and Meta Concept for a given word generated using the MetaMap [3] system.

- Filter words: The tagger automatically highlights several phrases to the human annotator, during the tagging process, based on the dictionary model. We capture explicitly the phrases which are removed by the human annotator during tagging process. These words are used to create a boolean flag feature which helps the system to eliminate some medical terms that are commonly disregarded by the emergency physicians.

*3.2.2 Sentence-level features.* Sentence-level features capture the context of the given word. These features focus on the previous and next words of the current word in the sentence. We defined the following sentence-level feature for our system.

- Previous and next word Part of Speech tags: These features help to identify the type of the current word. Similarly to the word-level POS, the sentence-level POS tags are generated from the MedPost [29].
- Next Negative and next positive words: This feature identifies the positive or negative sentiment words after the current word. The positive and negative sentiment word list are extracted based on the social media sentiment analysis [1]. The value of this feature is the actual positive or negative sentiment of the word.
- Previous and next negative word positions relative to the current word: This feature calculates how far the negative word is located from the current word. The negation word-list in this feature is based on the Negex [7] trigger word list. The value of this feature is the distance of the negative word from the current word.
- Word similarity: This feature compares the similarity of current word with the previous word. We used the word2vec [21] model for extracting this feature. The word2vec model was created based on 20,000 corrected radiology reports.
- Aggressive and Anatomy descriptors: These are boolean flags set to 1 if the anatomy or aggressive descriptors (from Radlex) are present in a 7-word window size of the current word (3 previous words + current word + 3 next words).
- High-flag, crit-flag, and non-crit flags: these flags check for the high critical, critical, and non-critical word presence in the 7-word window size. These dictionaries are created based on the manual annotations.

## 3.3 Information Extraction Module

In the Information extraction module, we identify the three different types of phrase from the radiology reports. These phrase are *high-critical*, *critical*, and *non-critical*. We used two types of phrase extraction model: 1) Dictionary based and 2) Machine learning.

*3.3.1 Dictionary based Model.* The dictionary based model is used for helping the human annotator to identify the possible phrases to annotate. This system uses the Radlex ontology to identify and highlight the phrases from the radiology report to be annotated. This model is a dictionary search-based model and cannot be used to identify the type of the phrase (For example, high-critical/critical). The interface for this model is given in Figure 2.

[1] https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107

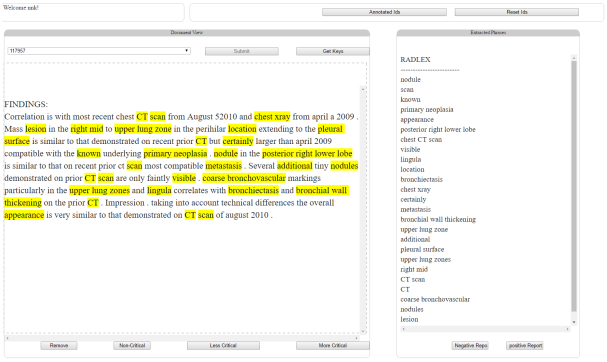


**Figure 2: The interface is designed for manual tagging of the radiology reports. The highlighted text is based on dictionary model.**

*3.3.2 Machine learning Model.* We have implemented two sequence learning models to evaluate the performance of the approach. The first is the CRF model which uses word level and sentence level features as discussed in Sections 3.2.1 and 3.2.2. This is a widely used machine learning model for identifying sequence labels. The CRF [16] model can be considered as a sequence-labeling version of logistic regression. It uses various feature functions (described in Sections 3.2.1 and 3.2.2) and learns weights to predict the sequence labels.

The model is trained to classify the phrases into three separate classes chosen after consulting with an emergency physician. Since the emergency physicians are primarily concerned with the immediate treatment of a patient's condition, it is necessary for the system to find medical phrases denoting conditions which have to be treated immediately. We use the classic BIO [6] model for labeling the training data. Prefix B-indicates the beginning part of the phrase and I-indicates the subsequent words. For example, B-Crit and I-Crit labels are used to indicate critical phrases, and the phrase 'heart is enlarged' is labeled as B-CRIT I-CRIT I-CRIT.

We also tried the structured Perceptron [8, 9] model to compare its performance with the CRF model. Structured Perceptrons is a version of Perceptron algorithms for sequence learning. Features used in this model are the same as the CRF model. The Structured Perceptron learns feature weights based on the auxiliary features extracted from each of the training samples and it uses the learned feature weights for predicting labels.

## 3.4 Document Classification Module

This module is used for classifying the radiology report into two classes namely critical and non-critical reports. The main purpose of classifying the report is for doctors to act quickly upon the critical reports. The document classification [15] system uses existing classification algorithms to classify the radiology reports. We have used those phrases that were extracted based on the Information Extraction module as an input to the classifiers and have compared the performance of the algorithms with 'bag of words' having tf-idf

[1] weights as features. Our purpose was to show the importance of the phrases and their extracted criticality level in classifying the reports.

*3.4.1 Algorithms and their parameters.* For report classification we used SGD (Stochastic Gradient descent), Random Forest and Linear SVM models. For each algorithm we compared the precision, recall and f1-score by using 'bag of words' having tf-idf [1] weights and then using critical level phrases extracted using the CRF model. The models were evaluated using 10-fold cross-validation and the average precision, recall and f1-score values of each model with the two types of features ('bag of words' having tf-idf weight and critical level phrases extracted using CRF model) were compared. For training the algorithms with extracted critical level phrases, we used a patient-level report vector, discussed in Section 3.4.2, along with some additional features such as word count and critical-level phrase count for each report.

*3.4.2 Document Matrix.* The Document Matrix is generated based on the output of the Machine learning model. It is essentially a vector for each of the patient reports where the columns represent the unique phrases extracted from all reports. This information can be used to quickly identify the condition of the patient from a collection of records. The level of criticality for each phrase is represented by a numeric value: a high-critical phrase is represented by +1, a critical-level phrase is represented by 0.5 and a non-critical level phrase is represented by -1.

## 3.5 Active Adaptive Learning Interface

Our Active Adaptive Learning Interface is the user interface which shows to the user the final phrases extracted and their criticality level. This interface can be used to edit the extracted phrases predicted by the model. The user can add/remove/change the criticality level of the phrases and the model is able to learn from the annotations of the given report in order to predict the phrases for the next report. This is achieved by including the predicted phrases and criticality levels as part of the binary level auxiliary features. This helps the system to provide higher weight to the observed word, based on corrected or previously predicted phrases. A sample screen-shot is given in Figure 3.

The interface is also able to provide the level of certainty for the phrases as well as the overall criticality level of the report. It provides a visual cue for the user, shown in a larger font-size, for those terms which are less certain. The user has the ability to edit the tag (criticality level) of the phrases. This extra information provides the user with the phrases which may have to be manually annotated. We focus only on the critical (high-critical/critical) level phrases and the OTHER type of phrases of the radiology report for providing the uncertainty levels. We omitted the uncertainty level for non-critical phrases to simplify the user's interaction and because these terms are usually not of interest of the emergency physicians. OTHER phrases are phrases which do not have any critical information (for example, medical phrases which are not tagged by the human annotator because they are not significance for judging the condition of the patient). OTHER phrases are phrases which are perceived by the system as having no information but potentially can have valuable information for the user. The system

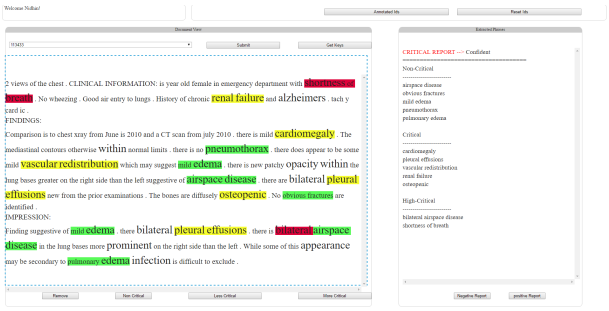

**Figure 3: Final Interface which highlights the information extracted from the radiology reports along with criticality levels for the phrases extracted. The overall document class (positive/negative) is shown at the upper right corner with the confidence level.**

determines a phrase as 'uncertain' based on the three cases given in the list below.

- Its high-critical prediction probability is at least 0.1 and predicted label is not high-critical.
- Critical prediction probability is at least 0.3 and predicted label is not critical.
- Predicted probability is less than 0.5 and predicted label is Other.

The system also provides the user with the overall criticality level of the report as well as the system predictions confident level. This can help doctors to identify emergency reports faster. A report is shown as a low-confidence prediction if the report class predicted distance is within one unit distance of the hyper plane. If the distance is more than one unit, it is predicted with high confidence. The distance score is negative for non-critical class and positive for the critical class.

# 4 RESULTS AND COMPARISON

In this section, we evaluated the proposed system performance. We divided this section into 3 parts: a) Results for the Word Segmentation module b) Results for the Machine Learning model c) Results for the Document classification.

## 4.1 Word Segmentation and Spell correction

The word segmentation module is used to segment the joined words present in the radiology reports. We used two methods to test our word segmentation module. Initially, we used a clean-text data set, which does not have any spelling errors, and we tested our model to check its accuracy. This provides us with an estimate of how many bogus word-segmentations are introduced, by the model, on clean text. For the second test, we created joined words (specifically radiology domain terms) and tested the system once again for the accuracy of segmentation.

For testing of the model with clean text, we used the text8 dataset [34] which contains over 3 million words. The text8 data is given to the algorithm for processing and we checked the number of words which are segmented by the model (ideally it should be 0). We obtained an accuracy of 98.9% on this data. This test was done

**Table 1: Accuracy of base and implemented models for joined word error correction.**

| | Accuracy (Our model) | Accuracy (Base model) |
|---|---|---|
| **Text 8 Dataset** | 98.90% | 98.90% |
| **Radlex random word combination 10k iterations** | 87.46% (bigram)<br>81.28% (trigram) | 42.58% (bigram)<br>25.69% (trigram) |

to make sure that the algorithm does not segment correct words present in real world documents.

For the second test, we created joined words from the words present in the Radlex ontology, chosen randomly and then combined together to create a joined word. The words chosen are medical words (not common words) in order to provide a better view of how well the system performs on uncommon words. We tested 2-word and 3-word combinations. The experiment was repeated for 10,000 iterations. We obtained an accuracy of 87.46% for 2-word combinations and 81.28% for 3-word combinations. This higher accuracy was obtained after adding the unigrams from the Google n-gram corpus for radiology terms (explained in detail on Section 3.1.1). Without adding the unigram radiology terms to the algorithm, the accuracy of 2-word combination was 42.58% and for 3-word combinations, it was 25.69%. This clearly shows that our model, with the addition of radiology terms, provides the best accuracy results. Table 1 shows the results in detail.

For evaluating the performance of the spelling correction algorithm, we used the text8 data set which has 3 million words. It was found that the algorithm produces an error rate of only 0.5%.

## 4.2 Machine learning Models

We used two sequence classifiers for our phrase extraction and criticality level identification. For each of the criticality levels, we used separate labels. For non-critical terms, we used B-NonCrit and I-NonCrit as the labels (Beginning word and subsequent word). Similarly, we used B-HighCrit, I-HighCrit, B-Crit, I-Crit respectively for high-critical and critical phrases. We used Conditional Random Field and Structured Perceptron as our two machine learning sequence classifiers.

*4.2.1 CRF.* We used the already existing fast implementation of Conditional Random Fields [16, 28] for our Model. The features used for the CRF are discussed in Section 3.2.1 and 3.2.2. We used l-bfgs [22] algorithm for the optimization. The coefficient values are dynamically calculated based on the training data.

We used 10-fold cross validation [26] on the training data. Since we do not check for inter-sentence parameters, the algorithm uses each sentence as training data. We obtained an average f1-score [30] of 0.75.

The average f1-score of non-critical, critical, and high-critical terms are 0.77, 0.70, and 0.78. The performance of critical terms is comparatively lower because of the higher uncertainty level for separating the high critical and critical terms (Figure 4). This boundary of separating critical and high-critical terms is heavily

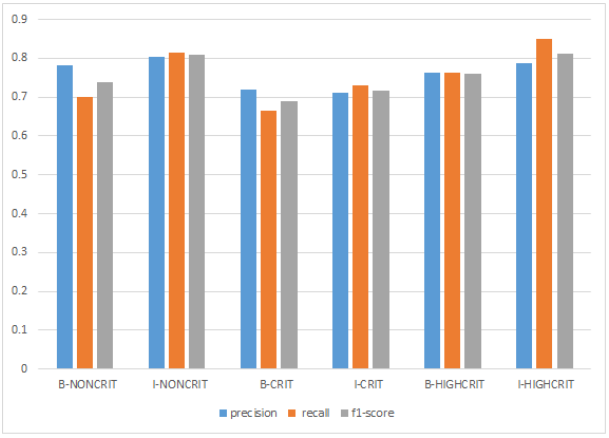


**Figure 4: Precision, Recall and f1-score of the CRF model for each of the different criticality level of extracted phrases.**

**Table 2: Confusion matrix of CRF model with various criticality level phrases. 'O' denotes 'Other' phrases which are irrelevant or are considered of no value to the doctors. 'B' and 'I' denotes the beginning and Intermediate words of the phrase.**

| | | Predicted labels | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | B-NONCRIT | I-NONCRIT | B-CRIT | I-CRIT | B-HIGHCRIT | I-HIGHCRIT | O |
| Actual Labels | B-NONCRIT | 63 | 4 | 2 | 1 | 2 | 0 | 21 |
| | I-NONCRIT | 1 | 61 | 1 | 2 | 0 | 1 | 9 |
| | B-CRIT | 2 | 0 | 35 | 1 | 5 | 1 | 8 |
| | I-CRIT | 0 | 1 | 1 | 27 | 0 | 3 | 4 |
| | B-HIGHCRIT | 2 | 0 | 4 | 0 | 40 | 2 | 2 |
| | I-HIGHCRIT | 0 | 1 | 0 | 1 | 1 | 32 | 2 |
| | O | 12 | 11 | 5 | 5 | 3 | 2 | 1354 |

dependent on the type of annotator (emergency physicians in this case). The confusion matrix of the model is given as in Table 2.

*4.2.2 Structured Perceptron.* The Structured Perceptron [8, 9] is also an existing model which is trained based on the auxiliary features which are used to train the CRF model. The Structured Perceptrons model works similarly to other sequence classifiers such as MEMM [19] and HMMs [4, 25]. However, Structured Perceptron's performance was less than the CRF model. It was able to provide an average f1-score of 0.72. The performance for non-critical terms was much worse than CRF model (0.68). The f1-score for non-critical terms was almost the same as the CRF model (0.76). And for high-critical terms, the f1-score was lower than the CRF model (0.76). The performance of structured Perceptron for the critical labels is shown in Figure 5.

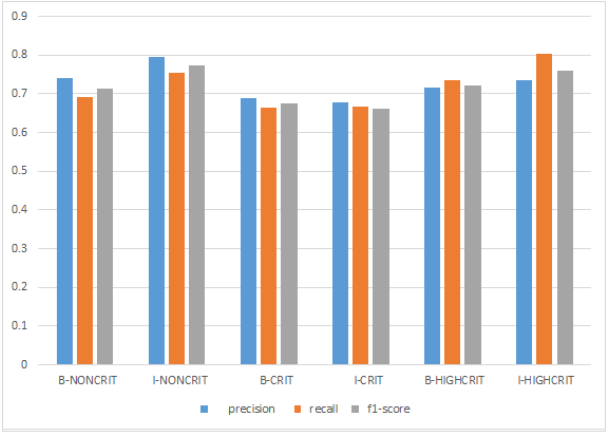


**Figure 5: Precision,Recall and f1 score of the Structured Perceptron model for each of the different criticality level phrases extracted**

*4.2.3 Comparing CRF and Structured Perceptron.* Both of these models are similar in performance. However, the CRF model performs better on average. The auxiliary features used for the training and prediction of sequence labels are the same. The CRF model is able to provide better recall than Structured Perceptron. Moreover, CRF provides the predicted probability values for the labels which can then be used for identifying the uncertainty of the predicted values. Evaluating the results for both CRF and Structured Perceptron by t-test, we obtain a p-value of 0.0549.

## 4.3 Feature Analysis

For our machine learning models, we have used two types of auxiliary feature:, word level, and sentence level features. In this section, we compare the models' performance based on the auxiliary features provided. For the sentence level features, we have segmented the performance graph into two parts, namely sentence level features, and flag level (or binary) features. The binary features are provided separately since the contribution of the binary features on the models' performance is significant.

As we compare the performance of the system based on the set of auxiliary features, the sentence and binary-level features provide a more significant contribution to the models' performance than do word-level features. One reason for this difference is that some of the word level features are inherently present in the sentence level features as well. For example, previous and next POS tags give similar contributions to assigning the POS tags of the current word. We assigned the current word POS tag contributes to the models' performance in special cases such as the beginning and end words of the sentence, and one-word sentences where there are no previous or next POS tags.

Binary features are part of the sentence-level feature-extraction module. These features are the main contributors to the Machine Learning model used in the active adaptive interface. These features are dynamically created based on the prior-tagged reports. For example, tagged phrases provided by humans during the training process are updated dynamically as the user uses the active/adaptive learning interface. These features create dictionaries based on the types (high-critical, critical and non-critical) of tagged critical phrases. These features help the model to identify medical terms which are critical or high-critical on most of the reports.

The combination of the three sets of features provides the best accuracy results for our model. The sentence-level features help to increase the recall value of our model while the word-level features are used to increase the precision of our model. The f1-score comparison for various features is provided in Table 3.

## 4.4 Inter Annotator Score

In order to compare our model to real-world human annotation performance we asked a second annotator to annotate the radiology reports and we then examined the consistency between the two sets of annotations.

The second annotator annotated 57 random reports out of the 253 reports tagged by the first annotator. For calculating the inter-annotator score, we used two methods. First, we used a 'soft' matching algorithm that only calculates the inter-annotator agreement on phrases which were annotated by both annotators. For the second method, we calculated the Precision, Recall, and f1-score of the second annotator on annotating the reports by keeping Annotator-1 as the gold standard. In both of these methods, we used the 57 reports annotated by the second annotator (Annotator-2).

The first evaluation method involves the calculation of the soft agreement score between annotators. The formula for the soft agreement score calculation is given in Equation 1.

$$Soft\ score\ =\ AVG\left(\sum_{i=1}^{57}\frac{W_i}{N_i}\right) \tag{1}$$

- $W_i$ = Number of words predicted by both annotators with same criticality level in report $i$.
- $N_i$ = Number of words predicted by both annotators in report $i$.

We obtained the soft agreement score of 71.47% on annotation. This proves that annotating a report and providing criticality levels to the phrases is a complicated task even for a human annotator who has ample domain knowledge. Moreover, reducing the annotation task to a 2-class system (critical/ non-critical) increased the inter annotation score to 85.01%. This experiment proves that the boundary of critical and high-critical can change based on the user's perception of each report. The confusion matrix for the soft score is shown in Table 4

The second evaluation method involves the training of the CRF model on the 200 reports that were not tagged by the second annotator. Once we trained the CRF model, we tested the model on the 57 reports tagged by the first annotator. We compared this result with the performance score obtained by asking the second annotator to tag the same 57 reports. The results are shown in Table 5. The CRF model gives similar performance to that of the human annotator but with higher precision. The performance dip in f1-score is due to the lower recall value, which would improved on an ongoing basis as the system acquires more data.

## 4.5 Document Classification

The radiology reports are classified into two classes, critical reports, and non-critical reports. The classification is based on the overall

**Table 3: Precision,Recall and f1 scores for CRF model based on various features used during training process.**

| | | B-NONCRIT | I-NONCRIT | B-CRIT | I-CRIT | B-HIGHCRIT | I-HIGHCRIT |
|---|---|---|---|---|---|---|---|
| **Word Level** | **precision** | 0.656 | 0.676 | 0.647 | 0.577 | 0.569 | 0.625 |
| | **recall** | 0.625 | 0.797 | 0.584 | 0.546 | 0.501 | 0.613 |
| | **f1-score** | 0.639 | 0.731 | 0.610 | 0.541 | 0.529 | 0.611 |
| **Sentence Level** | **precision** | 0.756 | 0.768 | 0.601 | 0.531 | 0.653 | 0.681 |
| | **recall** | 0.683 | 0.739 | 0.483 | 0.395 | 0.511 | 0.579 |
| | **f1-score** | 0.717 | 0.752 | 0.532 | 0.439 | 0.571 | 0.620 |
| **Binary Level** | **precision** | 0.706 | 0.735 | 0.738 | 0.704 | 0.712 | 0.755 |
| | **recall** | 0.593 | 0.769 | 0.591 | 0.698 | 0.679 | 0.790 |
| | **f1-score** | 0.643 | 0.750 | 0.653 | 0.694 | 0.692 | 0.769 |
| **Combined** | **precision** | 0.781 | 0.804 | 0.720 | 0.712 | 0.762 | 0.788 |
| | **recall** | 0.702 | 0.816 | 0.666 | 0.731 | 0.762 | 0.850 |
| | **f1-score** | 0.737 | 0.808 | 0.689 | 0.716 | 0.760 | 0.811 |

**Table 4: Confusion matrix for annotations done by second annotator on the radiology reports. Gold standard is based on the initial tagging done by the first annotator.**

| | | Predicted | | | | | |
|---|---|---|---|---|---|---|---|
| | | **B-NONCRIT** | **I-NONCRIT** | **B-CRIT** | **I-CRIT** | **B-HIGHCRIT** | **I-HIGHCRIT** |
| **Actual Labels** | **B-NONCRIT** | 129 | 7 | 6 | 0 | 13 | 0 |
| | **I-NONCRIT** | 7 | 126 | 0 | 5 | 1 | 10 |
| | **B-CRIT** | 11 | 1 | 16 | 8 | 33 | 5 |
| | **I-CRIT** | 1 | 12 | 0 | 15 | 3 | 17 |
| | **B-HIGHCRIT** | 8 | 0 | 8 | 0 | 95 | 9 |
| | **I-HIGHCRIT** | 0 | 2 | 0 | 8 | 7 | 75 |

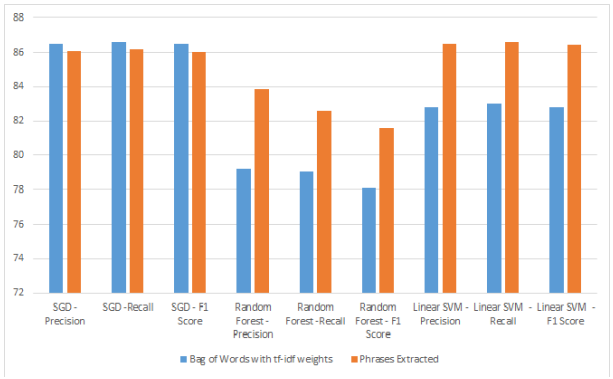


**Figure 6: Report classification results comparison using 'bag of words' having tf-idf weight as features and phrases extracted by the CRF model.**

report and is related to whether immediate action is required, on the patient in the emergency department. In order to analyze the relevance of the extracted phrases using the CRF model, we compared the classification accuracy of well known machine-learning algorithms using two methods. On the first trial, the reports are classified based on the 'bag of words' method having tf-idf weights assigned on those given in the report. In the second method, we used the phrases extracted using the CRF model along with the values assigned (-1 for non-critical phrases, 0.5 for critical phrases and 1 for high-critical phrases). We have used three separate machine learning algorithms (Linear Support Vector Machine, Random Forest and Stochastic Gradient Descent from the Sklearn library[2]) to compare the performance of each machine learning algorithm on these two types of feature. The comparison results are shown in Figure 6.

Comparing the results of the three algorithms on the two types of feature, we can see that the phrases extracted out-perform the 'bag of words' method having the tf-idf weights-based model on both the Random Forest [18] and Linear SVM [11]. Even on the SGD (Stochastic Gradient Descent) [2] the phrases extracted have similar performance to 'bag of words' having a tf-idf weights model. Also, the phrases extracted from the reports are comparatively much fewer than 'bag of words' having a tf-idf weights model.

Using the phrases extracted we were able to achieve an average f1-score of 86.42, in comparison to the average f1-score of 86.52 for 'bag of words' having a tf-idf weights model with SGD. These results demonstrate that the phrases extracted from the radiology reports are quite powerful features in classification.

Evaluating the statistical significance using the student t-test on the results, we have obtained a p-value of 9.43E-08 and 1.1E-05, respectively, for random forest and linear SVM and for 'bag of words' having tf-idf weights model and extracted phrase features. These results show that the 'extracted phrases' method performs better on classification of the report using these algorithms.

## 5 ERROR ANALYSIS

We have analyzed the misclassification errors for the CRF model which used the three level criticality levels for the extracted phrases. Upon analysis, the greatest misclassification occurs on classifying

[2]www.scikit-learn.org

**Table 5: Precision, Recall and f1-Score comparison between human annotator and CRF model.**

| | HUMAN ANNOTATOR | | | CRF MODEL | | |
|---|---|---|---|---|---|---|
| | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **B-NONCRIT** | 0.6825 | 0.6324 | 0.6565 | 0.8140 | 0.5122 | 0.6287 |
| **I-NONCRIT** | 0.7241 | 0.6632 | 0.6923 | 0.8947 | 0.6041 | 0.7212 |
| **B-CRIT** | 0.2712 | 0.1928 | 0.2254 | 0.4583 | 0.4074 | 0.4314 |
| **I-CRIT** | 0.2239 | 0.2500 | 0.2362 | 0.4643 | 0.2203 | 0.2989 |
| **B-HIGHCRIT** | 0.5220 | 0.7308 | 0.6090 | 0.6406 | 0.3228 | 0.4293 |
| **I-HIGHCRIT** | 0.5682 | 0.7353 | 0.6410 | 0.7692 | 0.4000 | 0.5263 |
| **Average** | 0.5703 | 0.5930 | 0.5759 | 0.7359 | 0.4564 | 0.5601 |

the non-critical phrases, which get classified as Other. These types of error are not a big concern in emergency-room practice since the doctors are mostly concerned about critical phrases. Even on manual tagging, depending on the report, some of the medical phrases may not be tagged by the doctor as non-critical. On analyzing the results, about 22% of the total non-critical phrases were predicted as 'Other' by the system. However, less than 2% of those terms which were actually non-critical were predicted as critical by the system.

Analyzing the critical phrases, the most common misclassification was, again, the classification of a 'critical' phrase as being 'Other'. However, the misclassification rate is lower compared to the non-critical phrases. The misclassification of critical phrases as Other is about 15%. However, on further analysis, it has been identified that the same phrase is misclassified in multiple reports which adds to the misclassification percentage. For example, the phrase 'Intrathoracic' is misclassified more than once, which adds to the misclassification rate even though only one phrase is misclassified. But this problem can be solved as we increase the amount of training data. As the doctors use the active adaptive learning interface through the on-line interface, these type of errors could be reduced considerably.

Finally, for high-critical terms, the most common errors are misclassification of the criticality level. About 8% of the high critical phrases are misclassified as critical phrases by the system. However, since the doctors are able to view both critical and high-critical phrases in the interface, along with the reports, these errors would not have a significant impact on the user experience.

## 6 CONCLUSIONS

We propose a system that performs extraction of medical phrases and their criticality level from free-text radiology reports, and classification of the whole report as being critical or not. As radiology reports are dictated by the radiologists and transformed into text, spelling and joined-word errors appear in the text, which we automatically correct, aiming to improve the accuracy of phrase extraction and classification. Information extraction from the radiology reports, in the form of medical phrases, is complex but provides valuable data, which can be further used in populating structured data bases for data mining tasks. The complexity of our task is due to the requirement of assigning the criticality level based on the textual context of the extracted phrases. The information extraction model, based on conditional random fields, extracts medical phrases and the associated criticality level (high-critical, critical and non-critical). The model is trained on a small corpus of reports labeled by two emergency physicians. We have demonstrated that our approach achieves performance that is comparable to the inter-annotator agreement. Using the extracted medical phrases as features, we address the report classification task that classifies entire radiology reports as critical or non-critical (i.e. whether an emergency physician needs to take immediate action on them). To allow the emergency physician user to efficiently inspect the extracted medical phrases and correct them if needed, we have built an adaptive active learning interface. Feedback provided by the user can be used for improving the performance of information extraction by on-line training.

## ACKNOWLEDGMENTS

We would like to extend our gratitude to Jessie Kang from the Department of Medicine for providing us annotation for 50+ radiology reports which helped us to compare the user agreement between annotations. We would also like to thank Palomino System Innovations for their guidance and support provided in this study. This research was funded by the Natural Sciences and Engineering Research Council of Canada.

## REFERENCES

[1] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.
[2] Shun-ichi Amari. 1993. Backpropagation and stochastic gradient descent method. *Neurocomputing* 5, 4-5 (1993), 185–196.
[3] Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 17.
[4] Jason Baldridge, Peter Clark, and Gokhan Tur. 2010. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. (2010).
[5] Olivier Bodenreider. 2004. The unified medical language system UMLS: integrating biomedical terminology. *Nucleic acids research* 32, suppl 1 (2004), D267–D270.
[6] Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003. A Simple Named Entity Extractor Using AdaBoost. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 152–155. DOI: http://dx.doi.org/10.3115/1119176.1119197
[7] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics* 34, 5 (2001), 301–310.
[8] Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 1–8.
[9] Hal Daumé III and Daniel Marcu. 2005. Learning as search optimization: Approximate large margin methods for structured prediction. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 169–176.

[10] Keith J Dreyer. 2014. Information theory entropy reduction program. (June 2014).

[11] Steve R Gunn and others. 1998. Support vector machines for classification and regression. *ISIS(Information Signals Images Systems) technical report* 14 (1998).

[12] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[13] Ferris M Hall. 2000. Language of the radiology report: primer for residents and wayward radiologists. *American Journal of Roentgenology* 175, 5 (2000), 1239–1242.

[14] Saeed Hassanpour and Curtis P Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine* 66 (2016), 29–39.

[15] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. 2007. Supervised machine learning: A review of classification techniques. (2007).

[16] John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML*, Vol. 1. 282–289.

[17] Curtis P Langlotz. 2006. RadLex: a new method for indexing online educational materials 1. *Radiographics* 26, 6 (2006), 1595–1597.

[18] Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[19] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation.. In *Icml*, Vol. 17. 591–598.

[20] Stéphane M Meystre, Julien Thibault, Shuying Shen, John F Hurdle, and Brett R South. 2010. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *Journal of the American Medical Informatics Association* 17, 5 (2010), 559–562.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[22] Jorge Nocedal. 1980. Updating quasi-Newton matrices with limited storage. *Mathematics of computation* 35, 151 (1980), 773–782.

[23] Peter Norvig. 2009. Natural language corpus data. *Beautiful Data* (2009), 219–242.

[24] Jon Patrick and Min Li. 2009. A cascade approach to extracting medication events. In *Australasian Language Technology Association Workshop November 28, 2009*. 99.

[25] Lawrence Rabiner and B Juang. 1986. An introduction to hidden Markov models. *ieee assp magazine* 3, 1 (1986), 4–16.

[26] Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2009. Cross-validation. In *Encyclopedia of database systems*. Springer, 532–538.

[27] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17, 5 (2010), 507–513.

[28] Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 134–141.

[29] L Smith, Thomas Rindflesch, W John Wilbur, and others. 2004. MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics* 20, 14 (2004), 2320–2321.

[30] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 1015–1021.

[31] Charles Sutton and Andrew McCallum. 2010. An introduction to conditional random fields. *arXiv preprint arXiv:1011.4088* (2010).

[32] Meliha Yetisgen-Yildiz, Martin L Gunn, Fei Xia, and Thomas H Payne. 2013. A text processing pipeline to extract recommendations from radiology reports. *Journal of biomedical informatics* 46, 2 (2013), 354–362.

[33] Wen-wai Yim, Tyler Denman, Sharon W Kwan, and Meliha Yetisgen. 2016. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Summits on Translational Science Proceedings* 2016 (2016), 455.

[34] Saizheng Zhang, Yuhuai Wu, Tong Che, Zhouhan Lin, Roland Memisevic, Ruslan Salakhutdinov, and Yoshua Bengio. 2016. Architectural Complexity Measures of Recurrent Neural Networks. *arXiv preprint arXiv:1602.08210* (2016).