Mixing Domain Rules with Machine Learning for Radiology Text Classification

Eamon Johnson · W. Christopher Baughman · Gultekin Ozsoyoglu

* Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, OH

† Department of Radiology
MetroHealth Medical Center
Cleveland, OH

ABSTRACT

In this work we compare and contrast proposed techniques and experiments in the subdomain of radiology text classification. In the past, systems have relied on two main approaches: rule-based or natural language processing plus machine learning (NLP/ML). Simple rule-based approaches have demonstrated great efficacy, at the high cost of requiring medical professionals to build and maintain. Complex NLP pipeline and ML approaches have also demonstrated efficacy, but rely heavily on computing professionals to build and maintain. To address this problem, we propose a hybrid classification mechanism for radiology text combining rules and NLP/ML that in our trials surpasses existing rule-based and NLP/ML techniques for the task of classifying incidental findings. We evaluate our approach against three approaches from related work using our own gold standard data set of 661 records. Our hybrid approach achieves a 13% F-measure gain over our prior rule-based approach[16] and a 4% F-measure improvement vs. a manual classification process in a hospital setting.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]

General Terms
Algorithms

Keywords
text mining, clinical text, radiology, incidental findings

1. INTRODUCTION

The practice of diagnostic radiology produces large volumes of free text interpretations of medical images. Secondary analysis of radiology reports has been a growing subdomain in medical natural language processing (NLP) over the last twenty years, driven by the successful application of traditional NLP tools to medical problems. Accurate extraction and classification of findings is central to NLP efforts in the radiology domain, yet no general solution to findings extraction has been developed. On the road to a general solution, specific information extraction and classification tasks have garnered much attention. Information extraction and classification for specific diseases and disorders are a very common type of task in radiology text processing [1, 2, 6, 8, 10, 11, 12, 24]. A second area that has received substantial attention is the classification of records containing incidental findings (unexpected or uncertain results) and the related problem of detecting recommendations for follow-up, which often stem from incidental findings. In this work, we focus on the problem of incidental findings detection.

Both rule-based and machine learning (ML) approaches have been shown to be effective for information extraction and classification tasks in the clinical text domain [7, 10, 15, 16, 17, 19, 20, 24, 25, 27, 31, 32, 33, 35]. Rule-based approaches allow experts to directly encode beliefs about data into decision trees which are then used for classification [7, 16]. The effectiveness of rule-based approaches implies that decision trees can create a binary classification of data based on a set of features. Further, assuming consistent distribution of data between the training set and the test set, the features used in rule-based binary classification of data should be learnable from the data, because the features must be present in the data or derivable from the data in order for the classification to be effective. A machine learning system ought to be able to compete with the performance of such a rule-based system for a given text extraction task, yet the characteristics of the data can confound even a sophisticated machine learning approach. Rule-based systems still persist for a number of reasons: simplicity of implementation, ease of understanding rationale for classification, and ability to include features underrepresented in the data.

In this work we discuss the compromises of rule-based and machine learning approaches and propose a mechanism for a hybrid approach to a specific clinical text classification task: detection of incidental findings in radiology reports.

2. BACKGROUND

2.1 Information Extraction in Diagnostic Radiology

Extraction and categorization of findings from radiology reports underpins a variety of clinically valuable secondary
analysis tasks, including document summarization and cohort identification. A more general history and background of information extraction from clinical text is available in [22]. Particularly to radiology, which relies on free text reports for communication between physicians, the development of reliable automated techniques for extraction of findings and certainty measurements could enable new quality management programs. For instance, the MetroHealth system of healthcare facilities in Cleveland, Ohio, has undertaken a quality improvement program to explicitly note the presence of incidental findings (unexpected or uncertain results) in radiology reports, which has increased capture and communication of follow-up recommendations from radiology to patients from 13% of exams in 2010 to 20% of exams in 2011 [9].

2.2 Incidental Findings and Recommendations for Follow-up

As opposed to critical findings, incidental findings often represent non-emergent conditions that will require further investigation or long-term monitoring. For example, at one Connecticut VA medical center, 52% of non-small cell lung cancers diagnosed between 2005 and 2010 were incidental findings [13]. Unfortunately, due to complexities in the delivery of medical care, clinicians often fail to inform patients of these incidental findings. Communication errors with regard to incidental findings are especially pronounced in emergency departments [28].

We draw a distinction here between two related but distinct problems in radiology text mining: (1) detection of incidental findings, and (2) detection of recommendations for follow-up. Incidental findings typically require additional imaging for further diagnosis, whereas in recommendations for follow-up the radiologist explicitly suggests that a patient receive additional imaging. Semantically, incidental findings are contextualized by the whole content of the report, whereas recommendations for additional imaging may be analyzed within an individual sentence. Further, incidental findings do not always require additional imaging, but may instead require long term monitoring. In our work, we analyze complete reports for signs of incidental findings, but we view the recommendation detection problem as similar enough that the techniques for recommendation detection should be effective for detection of most incidental findings.

Efforts to automate detection of incidental findings and follow-up recommendations have yielded promising results [7, 16, 31, 34] using supervised machine learning techniques in combination with NLP techniques and rule-based approaches. Detection of incidental findings can help to mitigate information loss in medical care settings, yet the predictive potential of fully identified and classified incidental findings [14] leads us to believe that contextualized finding extraction will provide clinical benefits beyond the scope of the immediate need to improve doctor-patient communication.

The LEXIMER system [3, 4, 5], now licensed to Nuance Communications, Inc., was used to classify diagnostic radiology reports containing recommendations for follow-up. Dreyer et al. reported 100% accuracy for classification of reports involving follow-up recommendations over a relatively small record set [5]. Whereas flawless accuracy in categorization is an unquestionably positive result, we show in Table 2 that the overwhelming percentage of cases do not include recommendations for follow-up, so many naive systems are able to achieve high accuracy without achieving a competitive F-measure.

2.3 Clinical Text Annotation

A major impediment to clinical natural language processing is the need for annotated corpora to train supervised classification machines. The task of extracting radiology findings is hampered, as with many other clinical NLP tasks, by the lack of training data. Training data is expensive to generate and once generated it is hard to validate, except by equally expensive cross-validation [23, 30]. Approaches to information extraction that rely on supervised learning do not scale well without methods to scale the learning process. In medical informatics, scaling the learning process is made difficult by institutional barriers to data sharing. For this work, we created a gold standard data set of records from a single hospital source (albeit a major trauma center), but we look forward to applying our techniques to larger, multi-institution radiology data sets in the future.

2.4 Rule-Based Systems

Rule-based systems for clinical text processing use collections of conditions on features to perform categorization. Rule-based systems are attractive because they are simple to generate and simple to understand. Yet such simple classification schemes can be prone to making broad, sweeping decisions, which can in turn increase the difficulty of balancing precision and recall. Two examples of rule-based approaches in the field of radiology text analysis are the system proposed by Dutta et al. [7] and the LHF algorithm described in our prior work [16]. The Dutta system uses rules on sentences and terms as compared with hand-made categorized term sets. The LHF algorithm uses hand-made regular expressions to express positive and negative concepts, and total weights are calculated for each sentence. Both approaches favor recall over precision—that is, they choose to decrease the false negative rate at the expense of the false positive rate. This is an understandable compromise given that false negatives directly create medical liability, whereas false positives can only create medical liability indirectly by wasting time.

An inherent weakness of the rule-based approaches is the inability to adapt to new data—they must instead be maintained via expert intervention. However, the fact that rule-based systems do not need to be trained allows them to represent conditions that either cannot be learned from the available training data, or which are difficult to discover or derive. The ability to classify unseen data based on expert-designed rules is a great strength, and in the medical text domain the designers of the rules are also typically the consumers of the system output. Additionally, since physicians must be involved in medical text classification systems as both author and audience, one can envision mechanisms to integrate the feedback of the audience into rule-based systems for long term maintenance of rules, as we suggest in [16].

2.5 NLP/ML Systems

Given the large variety of specialized clinical text processing pipelines available, researchers have the capability to quickly add an immense set of annotations to any given block of text. Automated examples in prior work have employed...
various methods to avoid the expense of expert manual labor, including unsupervised learning techniques [18], learning examples from the Unified Medical Language System (UMLS) [29], and using ontology-based measures of semantic similarity [21]. The sophisticated pipeline described by Yetisgen-Yildiz et al. [33] is specialized to classify sentences containing critical follow-up recommendations in radiology reports. The pipeline performs section-splitting, sentence-splitting, and then adds syntactic and semantic annotations to the sentences. The authors then compare the value of five feature sets for training a maximum entropy classifier: baseline (unigram), n-gram (2- and 3-), syntactic, knowledge-based (UMLS CUIs), and structural. The system was able to achieve an F-score of 0.717 using only the unigram features for sentence classification. The best F-score, 0.726, was achieved using the combination of unigram and syntactic features. This performance is shown in Table 2 in the YY-best column.

For the incidental findings classification problem, we observe that the unigram-based sentence classification approach is effective at identifying a large percentage of recommendations, but also adds to the false positive rate in an unacceptable manner. In this way, the result achieved by Yetisgen-Yildiz et al. effectively resembles the result achieved by Dutta et al., since both approaches take most of their benefit from term-matching.

3. OUR PROPOSED SYSTEM

Our system builds on state-of-the-art clinical natural language processing tools and incorporates multiple sources of information to construct a wide variety of features which can be used in machine learning for information extraction tasks. We then construct a classifier using a combination of rule-based and NLP/ML tools.

3.1 Our NLP Pipeline

The sequence of processing steps that constitute our NLP pipeline are illustrated in Figure 1. A typical radiology report consists of a number of sections, including, but not limited to, examination, clinical history, findings, and impression. Our section-splitter uses a set of 22 barrier tokens gathered via manual review of the data to divide the set of plain text records into sections. We use the cTAKES clinical NLP pipeline [26] to perform a set of standard NLP tasks: sentence splitting, sentence parsing, and part-of-speech (POS) tagging. The cTAKES pipeline also adds semantic annotations to text including indicators of confidence, polarity, uncertainty, use of generic terms, and use of conditional statements. The UMLS lookup features of cTAKES are able to add concept unique identifiers (CUIs) and SNOMED-CT annotations to the text, but we do not investigate the use of these features in this work, as they were shown to be of little value in [33]. After cTAKES processing is complete, we add additional features for use in machine learning by aggregating and propagating annotations made on words and sentences to their respective full-text reports.

3.2 Combining Rules and ML

Key to our approach is finding a satisfactory mechanism for mixing rules and learning processes. We consider three possibilities here:

1. ML on rules: where the verdict produced by rules is used as a feature for learning.
2. Rules on ML: where the verdict produced by a trained machine learning model is used as input for a rule set.
3. Complex composition: where rule-based decisions and learning steps are combined in a single decision tree.

3.2.1 ML on Rules

Using binary rule verdicts on records as features for machine learning is a simple task: the rule verdict may be added to the feature vector along with the other training and test inputs, and subjected to the same learning scheme. Rules producing continuous output may be harder to integrate into the learning process due to potential normalization issues between the training data and test data. With respect to the current text mining task, we confine our exploration to categorical verdicts.

Learning decision trees is a common task in supervised learning, and addition of rule-based decision nodes into the decision tree building process is straightforward. There does exist a possibility that a rule-based decision node could go unused if the tree building algorithm finds that the rule performs worse than other learned decision criteria for a given set of training data. The question then becomes how to force the model to include the rule-based decision node without compromising the tree’s decision-making capabilities. We will explore this question in future work.

3.2.2 Rules on ML

Just as binary rule verdicts can be added to feature vectors, so too can binary verdicts from machine learning be used as classification rules in hand-made decision trees. This approach may be fraught with the risk of unexpected classification depending on the level of understanding about the machine classifier held by the domain expert constructing the decision tree. If the domain expert fully understands the classifier’s decision process (for instance, in the case of a small decision tree model), then the ML classifier may be added to the tree at the point of best expected performance. For classifiers that are harder to understand intuitively, such as the n-dimensional hyperplane constructed by a support vector machine (SVM) classifier, the rationale of the classifier may be beyond simple explanation. Whether the working of the classifier can be easily explained or not, errors in classification may be attacked by design of further rules, as long as the designer is careful not to inject experiential bias by designing rules that fit specific conditions in the current data as opposed to rules that reflect general conditions from the domain.

3.2.3 Complex Composition

Complex composition indicates the combination of rule-decisions and ML-decisions into a single decision tree such that ML-training only occurs using the data that reaches each ML node. Unfortunately, the potential benefits of complex composition are difficult to weigh given our data and task. The foremost issue with complex combinations of rules and machine learning is the shrinkage of available data at each node in the tree. Given the class size imbalance in our gold standard set, with 48 positive records and 613 negative records, there will quickly become nodes where learning techniques are rendered ineffective due to scarcity of training examples.
3.3 Unigram Set Difference

We construct a novel training feature on our data by computing a set difference of the most significant unigrams between our two classes. We select the top $k=50$ normalized tokens (words) used in the positive and negative classes annotated in a training sample from our gold standard data set, and compute the set difference for each. Table 1 contains the listing of the most significant unigrams for each class in descending order of importance, as estimated by counting usage of each word. The negative list consists of words that resonate with intuition about situations that would generally not include incidental findings including views of bones, generalizations about soft tissue, and clear visualizations. The positive list reveals significant overlap with the categorical term list outlined by Dutta et al. in [7], which is to say that it includes terms reflecting advisory actions, uncertainty, and further tests.

4. EVALUATION

A thorough evaluation of our system will require curation of large clinical datasets overseen by multiple annotators – we plan to address strategies for accumulating large datasets in future work. Here, we use the task of extracting incidental findings as a proxy for complex clinical information extraction tasks in general. The problem of extracting incidental findings is substantially different from tasks involving identification of particular disease conditions, yet we believe it is a sufficiently nuanced problem to demonstrate the efficacy of our methods. In particular, extracting incidental findings requires interpretation of positive and negative findings, as well as understanding of the radiologist’s certainty and the context of the imaging.

We evaluate our system using a gold standard data set annotated by our co-author (W.C.B.). The data set consists of 661 radiology reports bearing binary annotation as to whether the record contains any incidental findings. This gold standard data set includes all radiology activity from one day at the MetroHealth system in Cleveland, Ohio. The base rate of incidental findings in our gold standard data set is 48 instances out of 661 total, or about 7%.

The performance of our proposed system can be directly compared to two incidental findings detection techniques using the same data set. First, we can compare our gold standard annotations to the annotations made by the radiologists who authored the reports. Second, we can compare our prior lexical-hints-with-feedback algorithm, which is essentially a rule-based approach when it is first initialized (i.e., without incorporating feedback). We also make a comparison to implementations of techniques for classification of recommendations for follow-up. In fairness to the designers of these techniques, we must recognize that the problem in our work – detection of incidental findings – differs from the problem that these techniques were designed to address: that is, detection of recommendations for follow-up.

4.1 Performance Measures

We employ a standard battery of performance measures as used in both medical and computing literature to assess the performance of classification tasks. The formulations for our measures are below, stated in terms of true positive (TP), true negative (TN), false positive (FP), and false negative (FN).
• Sensitivity, also known as recall: TP/(TP+FN).
• Specificity : TN/(TN+FP)
• Precision (positive predictive value): TP/(TP+FP)
• Accuracy: (TP+TN)/(TP+TN+FP+FN)
• F-measure: (2*TP)/(2*TP+FP+FN)

In the clinical setting, the main benefit of automated classification techniques is to reduce the amount of data that clinicians must review to make informed decisions. Recall is critical for reducing liability and precision is critical for minimizing time needed for secondary review. Efforts to maximize F-measure, which is the average of precision and recall, directly relate to balancing the sensitivity of classification against the volume of spurious results (false positives) that will require expert review after classification. Also important to note is the medical-legal liability implications of false negatives – the incorrect categorization of positive records as negative can create a situation where a healthcare provider has actionable information but does not communicate that information to the patient.

### 4.2 Related Systems

For purposes of comparison we present the best achieved measures of two related systems: Yetsigen-Yildiz et al.[33] and Dutta et al. [7], as drawn from their respective works, which are shown in Table 2 as YY-best and Dv3-best respectively. As discussed in Section 2.2, the problem of incidental findings detection strongly overlaps the problem of follow-up detection that is addressed by both of these systems, and we consider these to be the best examples in current research of systems comparable to ours. The YY-best system is an NLP/ML system which uses supervised learning on a corpus of sentences extracted from radiology reports to perform classification of sentences containing recommendations for follow-up. The Dv3-best system is a rule-based system which uses a combination of word search, categorical rules, and negation handling to perform classification of sentences containing recommendations for follow-up.

**Table 2:**

Performance of Recommendation Detection Systems

<table>
<thead>
<tr>
<th></th>
<th>YY-best</th>
<th>Dv3-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>100</td>
<td>70</td>
</tr>
<tr>
<td>TN</td>
<td>18583</td>
<td>1488</td>
</tr>
<tr>
<td>FP</td>
<td>51</td>
<td>75</td>
</tr>
<tr>
<td>FN</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>sensitivity (recall)</td>
<td>0.885</td>
<td>0.972</td>
</tr>
<tr>
<td>specificity</td>
<td>0.997</td>
<td>0.952</td>
</tr>
<tr>
<td>precision (PPV)</td>
<td>0.662</td>
<td>0.483</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.997</td>
<td>0.953</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.758</td>
<td>0.645</td>
</tr>
</tbody>
</table>

### 4.3 Comparative Systems

We present a set of eight classifiers applied to our data. These include one manual process at a hospital, five rule-based systems, and two systems using machine learning, of which one also uses rule-based features. The first six are presented for establishment of baseline performance of classification techniques on our data set, reflected in the first six columns of Table 3. The final two columns of the table show the performance of two new classifiers produced for this work. When reading the table, note that the ML classifiers use a randomized 70/30 train/test split, so the counts given are only for the classification of the test data, which is 30% of the total. Due to the class size imbalance we applied the randomized split to the classes separately so as to ensure that the class balance within the train and test sets was the same as the class balance in the data as a whole.

- The Hospital column indicates the performance of a manual incidental finding tagging process in daily operations at the facility where the records were created. The Hospital process did not create any false positives, but also has low recall, which in real-world terms indicates lost information. Note that these annotations were made after the implementation of a concerted quality improvement program to increase capture of incidental findings.
- The LHF column represents the Lexical Hints with Feedback algorithm from our prior work [16], which offers substantial gains over the hospital process but suffers from low precision due to a high false positive rate.
- Dv1 and Dv2 are our implementations of the first and second iterations of the algorithm from [7]. The authors do not provide full details of the mechanism for the third iteration, presented as Dv3-best in Table 2. Both Dv1 and Dv2 are rule-based algorithms which use categorized word search to classify records. We note that Dv1 has a near-perfect true positive rate on our data, and that adding the word “probably” to the keyword list would have brought recall to 100%. Of course, high recall comes at the cost ofprecision (positive predictive value), which in turn negatively impacts F-measure. The Dv2 algorithm sacrifices recall in favor of precision by adding a requirement that a sentence contain terms from at least two categories, a move which greatly increases F-measure.
- UC1 represents a simple rule: if any semantic annotation added to the record by cTAKES contains the attribute/value combination “uncertainty=1”, then the record is classified as positive. Importantly, this rule uses no word list or human intervention, but displays performance similar to Dv1 which is driven by a hand-curated word list.
- The ME1 system is a naive maximum entropy classifier using normalized unigram composed into weighted feature vectors and a 70/30 train/test ratio. ME1 achieves high specificity and accuracy, but this is not significant due to the imbalance in the positive and negative class sizes in our data. If we were classifying sentences containing follow-up, we would expect this approach to approximate the baseline/unigram accuracy from [33], but the low F-measure of this result demonstrates the increased difficulty of classification for the incidental findings detection problem as opposed to the follow-up recommendation detection problem.
The Rule2 column represents a decision tree classification using only a single two-part rule, which is applied only to the Impression and Findings sections of each record: (1) positive records must contain a word from the Dutta et al list of advisory words (“recommend”, “follow-up”, etc.), and (2) the record must contain a procedure mention, as identified by the semantic annotation output of our NLP pipeline. This rule resembles the Dv2 algorithm, but differs in that it uses semantic tags.

The Hybrid column represents a hybrid rule and machine learning approach, where the output of the rule-based classification is added to the feature vector of the machine learning process, as described in Section 3.2.1. As in the Rule2 approach, only the Impression and Findings section of each record are considered. This system uses a decision tree learning algorithm over the following features: normalized unigrams, positive and negative unigrams from our set difference list in Section 3.3, binary verdicts from rule systems (Dv1, Dv2, LHF), and the semantic annotations output by our NLP pipeline. A 70/30 train/test ratio was used.

4.4 Discussion

The results of our classification experiments show our Hybrid approach yields 13% better F-measure than our prior rule-based system (LHF), and 4% better F-measure than the manual process in place in the hospital. These results show the value of integrating NLP pipelines with rule sets and semantic tags into classification tasks. These results are encouraging, and in the future we hope to support the validity of our findings through application to larger and more heterogeneous data sets.

5. CONCLUSIONS AND FUTURE WORK

The current state-of-the-art in medical natural language processing is limited by a lack of annotated data. Methods to learn about the text from the text rather than from expert annotators are attractive because of the extreme cost of expert annotation, yet these methods ultimately must be validated against expert opinion.

Our proposed system represents a hybrid rule and NLP/ML approach to detection of incidental findings, combining multiple sources of data with a limited rule set based in domain knowledge. We have shown that we are able to accomplish a complex information extraction task using a small set of domain-specific rules, with better performance than a solely rule-based or machine learning approach.

6. REFERENCES


<table>
<thead>
<tr>
<th>Hospital</th>
<th>manual</th>
<th>LHF</th>
<th>Dv1</th>
<th>Dv2</th>
<th>UC1</th>
<th>ME1</th>
<th>Rule2</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>23</td>
<td>39</td>
<td>46</td>
<td>32</td>
<td>38</td>
<td>7</td>
<td>31</td>
<td>10</td>
</tr>
<tr>
<td>TN</td>
<td>613</td>
<td>562</td>
<td>328</td>
<td>549</td>
<td>419</td>
<td>303</td>
<td>598</td>
<td>179</td>
</tr>
<tr>
<td>FP</td>
<td>0</td>
<td>51</td>
<td>285</td>
<td>64</td>
<td>194</td>
<td>2</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>FN</td>
<td>25</td>
<td>9</td>
<td>2</td>
<td>16</td>
<td>10</td>
<td>19</td>
<td>17</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sensitivity (recall)</th>
<th>0.479</th>
<th>0.813</th>
<th>0.958</th>
<th>0.667</th>
<th>0.792</th>
<th>0.269</th>
<th>0.646</th>
<th>0.714</th>
</tr>
</thead>
<tbody>
<tr>
<td>specificity</td>
<td>1.000</td>
<td>0.917</td>
<td>0.535</td>
<td>0.896</td>
<td>0.684</td>
<td>0.993</td>
<td>0.976</td>
<td>0.972</td>
</tr>
<tr>
<td>precision (PPV)</td>
<td>1.000</td>
<td>0.433</td>
<td>0.139</td>
<td>0.333</td>
<td>0.164</td>
<td>0.778</td>
<td>0.674</td>
<td>0.666</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.962</td>
<td>0.909</td>
<td>0.566</td>
<td>0.879</td>
<td>0.691</td>
<td>0.937</td>
<td>0.952</td>
<td>0.954</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.648</td>
<td>0.565</td>
<td>0.243</td>
<td>0.444</td>
<td>0.271</td>
<td>0.400</td>
<td>0.660</td>
<td>0.689</td>
</tr>
</tbody>
</table>


