Sharing is Caring: Exploring machine learning-enabled methods for regional medical imaging exchange using procedure metadata

Salaar Liaqat, BSc¹, Joanna Pineda, BSc¹, Jeevaa Velayutham, BSc¹, Allen Lee, BSc¹, Joshua Reicher, MD², Jason Nagels³, Marzyeh Ghassemi, PhD¹,4, Benjamin Fine, MD¹,5
¹University of Toronto, Toronto, Ontario Canada;
²Stanford Health Care, Stanford, California, USA;
³Hospital Diagnostic Imaging Repository Services, Toronto, Ontario, Canada;
⁴CIFAR AI Chair at Vector Institute, Toronto, Canada;
⁵Trillium Health Partners, Toronto, Ontario, Canada

Abstract

Seamless sharing between imaging facilities of medical images obtained on the same patient is crucial in providing accurate and efficient care to patients. However, the terminology used to describe semantically similar examinations can vary widely between facilities. Current practice is manual table-based mapping to a standard terminology, which has substantial potential for mislabelled and missing examinations. In this work, we establish several baseline methods for automating the mapping of radiology imaging procedure descriptions to a SNOMED CT based standard terminology. Our best performing baseline, consisting of a bag of words representation and shallow neural network, achieved 96.3% accuracy. In addition, we explore an unsupervised clustering method that explores relevancy matching without the need for an intervening standard. Lastly, we make the procedure name dataset used in this work available to encourage extension of this application.

Introduction

Medical imaging provided by healthcare practitioners plays a central role in modern medical diagnosis and treatment. Ideally, interpretation of any exam should incorporate all relevant prior exams for a patient from all imaging institutions. In practice, this is challenging because exam metadata - specifically, procedure descriptors such as names, codes, body parts or modalities (referred to as “procedure names” in this paper) - are developed locally without a view to generalizability or interoperability across imaging facilities. This presents a unique challenge of learning which procedure names are semantically identical between sites, but are represented by different (but similar) strings of characters (e.g. “Chest XR” vs “X-ray Chest”). In the following sections, we describe the current approach to cross-facility Diagnostic Image (DI) exchange in Ontario, Canada using Diagnostic Imaging Repositories (DIRs) and explore machine learning-based approaches to augment and automate accurate cross-site image sharing.

The primary purpose of a DIR (akin to Health Information Exchange) is to host and provide a patient’s longitudinal imaging records to hospitals and clinics in a specific geographic region, independent of the site at which images are acquired. There are 4 such DIRs in the province of Ontario (pop. 14.5 M), centrally storing diagnostic images for all hospitals (excludes most private clinics). Currently, when facilities generate new DI exams they are assigned a procedure name using local codes and procedure descriptions specific to their institution; a copy of the examination is then archived to the DIR for storage and distribution to other hospitals and clinics. In an effort to facilitate accurate imaging exchange, the Hospital Diagnostic Imaging Repository Services (HDIRS) created a set of regional codes based on a standardized collection of medical terms used for documentation and reporting (SNOMED CT). To implement this standard, each member site must manually map their local procedure codes to the standard DIR terminology. When a site has thousands of procedures, it becomes very time consuming to find the corresponding regional procedure for each local procedure. Most importantly, managing changes and maintaining mapping accuracy is also done manually and often neglected. This can lead to a disorganized and unmaintained DIR with orphan images, which hinders the potential of image sharing between sites and can produce functional gaps in the longitudinal patient record.

In this work, we use Natural Language Processing (NLP) techniques to find an effective representation of procedure names, and Machine Learning (ML) methods to learn an effective mapping to the standardized terminology defined by HDIRS, allowing for automation of the error-prone manual mapping process currently in use. We evaluate these
methods, and identify the best representation and model combination in mapping local procedure name terminologies to regionally defined ontology. We then extensively evaluate various aspects of this state-of-the-art method, specifically demonstrating where the model succeeds and fails in order to provide explainability to users. In our last section, we begin to explore the applications of NLP methods to create sets of semantically meaningful clusters. This technique could enable efficient imaging exchange without the need for an intervening standard terminology, which is manually created by the DIR and independently repeated by other central imaging repositories across the globe.

Our specific contributions are as follows:

- Describe and release an Ontario-wide procedure code dataset for public-use
- Develop competitive baselines for automating the mapping of site-specific study description to regional codes
- Apply extensive evaluation to the best performing method
- Describe an unsupervised clustering method to enable procedure retrieval without a standard terminology

Related Work

Huang et al., Irvin et al., Mullenbach et al. and Singh et al. have explored NLP and machine learning techniques for learning mappings of text to labels or codes in medicine. These methods are applied using doctor’s notes, and other medical reports but do not specifically address procedure names, which have a unique language structure. Nguyen et al. and Patel et al. have demonstrated that a combination of natural language processing and machine learning methods are robust to nuances such as spelling errors, and are preferred in clinical settings over rule-based approaches.

In the medical imaging domain, Fidahussein et al. used Apache’s OpenNLP to develop supervised models to automatically map laboratory procedure names from local terminologies to a LOINC terminology, another standard medical ontology. Using the Maxent and Lucene models from OpenNLP, they achieved accuracy of 79% and 72%, respectively. Similarly, Intelligent Mapper (IM) is a tool to automatically determine the LOINC code for local terms. IM uses rules-based matching to find the correct label, and achieved a test accuracy of 72.87% on a single unseen hospital dataset.

Previous work has explored the mapping of various standards of medical metadata to a standard. Although achieving relatively high accuracy, these methods are rule based or applied to small datasets. In this paper, we use a larger dataset, obtained from over 90 hospitals, for training of a machine learning-based model and demonstrate notably higher accuracy.

Dataset

In the following section, we describe the dataset utilized in this study, which represents the local procedure names for over 90 imaging facilities in Ontario (N = 168,959). Note that local site information was only available for 20 sites (N = 40,923), which was removed from the publicly-released dataset. This dataset has been made publicly available by HDIRS to encourage extensions of this work and facilitate the movement towards shared medical information. The dataset can be found at: https://github.com/saliaqat/ontario-local-to-snomedct-mapping-dataset

1 This dataset has been made publicly available by HDIRS to encourage extensions of this work and facilitate the movement towards shared medical information. The dataset can be found at: https://github.com/saliaqat/ontario-local-to-snomedct-mapping-dataset
Table 1: Three example image metadata from the dataset. The first seven columns are textual descriptions of each image from the RIS and PACS. These values can be empty. The last column is the regional label defined for specific image/procedure and acts as our class for classification.

<table>
<thead>
<tr>
<th>RIS</th>
<th>PACS</th>
<th>code</th>
<th>description</th>
<th>code</th>
<th>procedure</th>
<th>study</th>
<th>body part</th>
<th>modality</th>
<th>regional label</th>
</tr>
</thead>
<tbody>
<tr>
<td>uskidwo</td>
<td>ultrasound kidney w/o rad</td>
<td>uskidwo</td>
<td>-</td>
<td>us</td>
<td>us</td>
<td>gu</td>
<td>us</td>
<td></td>
<td>101557</td>
</tr>
<tr>
<td>pehlal</td>
<td>pelvis ap hip lt ap/lat</td>
<td>pehlal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>hiplwex dx</td>
<td></td>
<td></td>
<td>101248</td>
</tr>
<tr>
<td>hipl1</td>
<td>hip left 1 view</td>
<td>hipl1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>hiplwex dx</td>
<td></td>
<td></td>
<td>101178</td>
</tr>
</tbody>
</table>

should not occur in theory, but do appear in practice. This mapping process occurs every time a new local site joins a DIR, or when a site changes their Hospital Information System (HIS). The Ontario-wide dataset provided has 168,959 manually mapped procedure names from approximately 90 sites. The first seven columns represent fields from the local RIS and PACS. The eighth column of each row has a manually-mapped numerical regional code defined, which maps the local terminology to the regionally defined procedure. This dataset contains solely the mapping of procedure codes and thus, does not contain any personal health information. Three examples of the data are shown in Table 1.

Characteristics The dataset has defined a standard set of 1,855 regional codes (with corresponding procedure names) to which all sites manually map local procedure data. We use these regional codes as prediction classes for our models. Reflecting real world variation in many to 1 mapping of local to procedure names, there is a predicted class imbalance. The predicted class frequency in the dataset ranges from 0 to 3740, with the median occurrence being 18. Furthermore approximately 75% of the data is made up of 20% of the classes. This imbalance, presented in Figure 1a and b, shows that low occurrence classes are predominant in the dataset, and correct prediction of these classes is a necessity for a strong model.

There are seven feature columns, consisting of text describing the procedure. We calculated the amount of information presented in each row by counting the number of words in a row, shown in Figure 1c. We found that a row could contain anywhere from 3 to 33 words, with an average near 15. Although this can be viewed as a small number of words in the NLP domain, most words in the dataset consist of keywords and virtually no stop-words. This produces an information rich and dense text dataset. However, because the RIS and PACS procedure data elements were inconsistently entered across sites, the input matrix was sparse; no single input data element could be relied upon. These data characteristics motivated our data preparation as described below.

Figure 1: a) Distribution of class frequency in the dataset b) Cumulative distribution of class frequency in dataset c) Frequency of number of words in each row in the data

Data Processing

In the following subsection we describe our preprocessing steps to represent the data.

Data Preparation Based on the domain knowledge of our team, we identified that word order was not a significant indication of the class. For instance, “chest x-ray” is synonymous to “x-ray chest”. With this observation, and the ones mentioned above, we merged all the columns to form a single string for a procedure. We then tokenized this string based on whitespace, using NLTK[^17], leaving us with a tokenized vector of words for each row, and a corresponding
regional label. This form of data is similar to short texts as discussed by Zuo et al.\(^\text{18}\) and served as the basis for the representations for our models.

**Data Representations** We first looked at a bag-of-words (BOW) representation\(^\text{19}\). BOW creates a vector where each element corresponds to a word in the dataset. For each word in an input, the corresponding element in the vector is incremented. As a result, which words occur and their counts are preserved, however, the order of words is lost. We also employed TF-IDF, which has the same characteristics as BOW; however, rather than recording the counts of the words, a weight is recorded corresponding to the frequency of the term in the input, which is then inversely proportionate to the frequency of the term in the entire dataset. These two representations encode every word, but lose the order which aligns well with our observations that each word carried relevant information, but the order was not necessarily important.

We also investigated document embeddings to leverage data sparsity and domain specific information\(^\text{20}\). We trained a Doc2Vec model, which was created by Quoc and Mikolov\(^\text{21}\). A Doc2Vec model analyzes the structure of a sentence or document, and maps it to a vector space. As a result, similarly structured sentences would map to similar vector spaces. Document embeddings is a new field, which has emerged as an extension of word embeddings which has many models, optimized for various tasks and domains. Because word embeddings are commonly used in machine learning, we explore the use of a pre-trained Word2Vec model. In the medical domain, a model called PubMed\(^\text{22}\) was trained on over 29 million medical journals and should be able to distinguish better between medical topics. We also use this word embedding to embed every word in a row. We then transform the word embedding to a document embedding by taking the average of every word embedding for a row. This method of representing short texts using word embeddings was presented and validated by Liu et al.\(^\text{23}\).

Lastly, we observed that many words in the dataset were semantically identical, but had different spellings. This was caused by the use of shorthand, alternate words, or misspellings. This caused our BOW and TF-IDF representation to become bloated, and would cause inefficient models. To alleviate this, we used an autoencoder on BOW to compress our representation with the goal to map semantically identical words to the same element in the representation vector. An autoencoder attempts to recreate the input as an output using a neural network, using a bottleneck layer in the center. After training, this bottleneck layer would encode the input as a compressed representation\(^\text{24}\).

**Models**

**Procedure Name Classification (Supervised)** We first developed multi-class logistic regression and random forest models\(^\text{19}\). Both these models tend to perform competitively when compared to larger, more complicated models and work well with real-world data. Furthermore, these models were expected to work well with simpler input data like BOW and TF-IDF, as the operations performed by these models work well on less complicated data. We also developed a naive Bayes model\(^\text{19}\) due to its historic success on textual categorization\(^\text{25}\) and medical diagnostic\(^\text{26}\). Furthermore, naive bayes uses distributions to form models of the data, which falls into line with how embeddings map similar data to similar vector spaces. Thus, we hoped naive Bayes could leverage the domain knowledge encoded by word embeddings.

We also constructed a fully connected, feed-forward, neural network in Keras\(^\text{27}\). The neural network (NN) consisted of a single hidden layers, with 4096 units. The layers used ReLu activation functions with the output layer using a softmax to predict a single class out of the possible 1,855 classes. The model uses the Adam optimizer to minimize the categorical cross-entropy loss\(^\text{28}\).

**Training** For supervised training we randomly split the data from all sites into training, validation and testing sets (75%, 12.5%, 12.5%), and formed the representations using the training set. We then used the representation algorithms developed for the training set to transform the validation and testing sets.

**Procedure Name Clustering (Unsupervised)** Lastly, we explored the potential of using unsupervised clustering methods to form clusters based on the inputs. With this model, we considered the possibility that a regional standard for procedure codes didn’t exists. This is particularly applicable, considering we are releasing our dataset. The classes defined in our dataset may have limited use to researchers in different geographical regions. In that case, researchers
Table 2: Evaluation of all representation and model combinations. A bag of words representation with a neural network achieved the best accuracy. The model also achieved among the best precision, recall and F1 scores.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Macro Precision (%)</th>
<th>Micro Precision (%)</th>
<th>Macro Recall (%)</th>
<th>Micro Recall (%)</th>
<th>Macro F1 Score (%)</th>
<th>Micro F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc2Vec</td>
<td>Logistic Regression</td>
<td>2.7</td>
<td>0.1</td>
<td>0.4</td>
<td>0.1</td>
<td>2.7</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>0.6</td>
<td>0.1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.6</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>71.6</td>
<td>43.3</td>
<td>54.4</td>
<td>57.1</td>
<td>71.6</td>
<td>48.3</td>
<td>61.3</td>
</tr>
<tr>
<td>PubMed</td>
<td>Logistic Regression</td>
<td>83.8</td>
<td>49.0</td>
<td>72.6</td>
<td>56.3</td>
<td>83.8</td>
<td>51.8</td>
<td>77.4</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>86.9</td>
<td>64.5</td>
<td>78.2</td>
<td>73.9</td>
<td>86.9</td>
<td>68.0</td>
<td>81.9</td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>74.0</td>
<td>57.5</td>
<td>61.3</td>
<td>68.4</td>
<td>74.0</td>
<td>61.0</td>
<td>66.0</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>94.1</td>
<td>78.2</td>
<td>90.1</td>
<td>83.7</td>
<td>94.1</td>
<td>80.3</td>
<td>91.9</td>
</tr>
<tr>
<td>Autoencoder</td>
<td>Logistic Regression</td>
<td>92.6</td>
<td>74.8</td>
<td>87.7</td>
<td>81.4</td>
<td>92.6</td>
<td>77.3</td>
<td>89.8</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>89.8</td>
<td>68.1</td>
<td>82.9</td>
<td>76.7</td>
<td>89.8</td>
<td>71.3</td>
<td>85.8</td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>82.4</td>
<td>64.2</td>
<td>74.7</td>
<td>70.6</td>
<td>82.4</td>
<td>66.5</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>91.0</td>
<td>68.6</td>
<td>85.2</td>
<td>76.4</td>
<td>91.0</td>
<td>71.5</td>
<td>87.6</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>Logistic Regression</td>
<td>72.9</td>
<td>60.5</td>
<td>85.3</td>
<td>69.8</td>
<td>72.9</td>
<td>63.8</td>
<td>63.7</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>79.2</td>
<td>64.0</td>
<td>66.6</td>
<td>77.7</td>
<td>79.2</td>
<td>69.1</td>
<td>71.6</td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>79.0</td>
<td>66.5</td>
<td>66.0</td>
<td>81.1</td>
<td>79.0</td>
<td>72.1</td>
<td>71.2</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>95.8</td>
<td>83.2</td>
<td>92.7</td>
<td>86.5</td>
<td>95.8</td>
<td>84.6</td>
<td>94.2</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>Logistic Regression</td>
<td>95.1</td>
<td>75.5</td>
<td>91.1</td>
<td>79.7</td>
<td>95.1</td>
<td>77.3</td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>92.2</td>
<td>70.1</td>
<td>86.1</td>
<td>76.7</td>
<td>92.2</td>
<td>72.9</td>
<td>88.9</td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>88.1</td>
<td>67.9</td>
<td>81.4</td>
<td>78.5</td>
<td>88.1</td>
<td>71.5</td>
<td>84.1</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>96.3</td>
<td>84.4</td>
<td>93.5</td>
<td>87.5</td>
<td>96.3</td>
<td>85.7</td>
<td>94.8</td>
</tr>
</tbody>
</table>

may be interested in deriving new classes from the data. We conduct a preliminary analysis to demonstrate the feasibility of this approach. We use a k-means unsupervised clustering method. However, since we did not utilize the labels for this model, we could not objectively evaluate the model, and resorted to a qualitative evaluation.

Results

Quantitative model evaluation We evaluate accuracy, macro/micro precision, recall and F1 score for each supervised model, trained with each representation. The macro scores treat each class with equal weight, giving a score that represents all classes in the dataset. This score would be lower if some classes were inaccurately represented by our model. Conversely, micro scores give each class a weight corresponding to occurrence frequency, giving a score that is more representative of how the model would perform with the dataset. To ensure that the model is not discriminating against certain classes, and is performing well overall, we want both the micro and macro scores to be similarly high. Results of this evaluation are shown in Table 2.

Generalization check Since data from one imaging site can vary in terms of terminology and the distribution of labels, observing accuracy for an unseen site would demonstrate the generalizability of our best performing model. With originating information on 20 sites in our dataset, we evaluated model accuracy by training on \( x \) random sites without replacement and testing on the remaining 20 – \( x \) sites, for \( x = \{1, 2, \ldots, 19\} \). This evaluation was run 15 times to obtain a distribution of accuracy scores, shown in Figure 2. Any data point where the originating site was unknown was kept out for this evaluation.

Class Evaluation An ideal procedure name mapping model would perform accurately for all classes. However, the previously noted widely skewed class imbalance introduces challenges for evaluating real world model performance. Due to the high number of classes, looking at all classes in depth was infeasible. Instead, we form a confusion matrix for each class, which would show a high-level view of the model performance. We formed the confusion matrix such that the top left elements would be the highest occurrence classes and moving down the matrix would contain fewer occurrence classes. This confusion matrix can be found in Figure 3.

Aggregate Class Evaluation Beyond comparing all classes between themselves, which results in a 1,855 by 1,855 confusion matrix, we can merge classes based on similarities. For medical procedures, we can merge classes that have
Figure 2: Model accuracy when trained on data from x hospitals and tested on data from remaining hospitals. Training and testing hospitals were chosen randomly without replacement, and run 15 times.

the same modality. For instance, all procedures that involve using a CT scan would fall into the same class, and we can evaluate our models ability to select a procedure with the same modality. We can also use the body part that a procedure examines as a common feature to group classes. These confusion matrices were also formed with the most popular classes in the top left and rarest in the bottom right. We present these confusion matrices in Figure 4.

Qualitative evaluation of unsupervised clustering

As the regional procedure codes were defined by human experts at HDIRS, we explored inferring labels without a human-defined standard terminology using unsupervised clustering methods. Since semantically identical procedures will have syntactically similar text, we expect the model to separate the data based on the procedure. This is lightly enforced by defining the model to contain approximately 1,500 clusters. We used k-means clustering implemented with scikit-learn and the TF-IDF representation described above to create and train the model. We show the top 10 keywords for three example clusters in Table 3.

Discussion

Quantitative model evaluation

We achieved competitive performance with all our representations demonstrating that were capable in finding representations for the input. However, only the neural network model was able to understand the Doc2Vec representation. The best performing representation was BOW. This aligns with our characterization of the data, where we highlighted that word order was not an important features but ensuring that every word was recorded was. We hypothesize this is why BOW was such an effective representation.

The overall method that performed best was BOW with a NN, achieving 96.3% accuracy and slightly lower precision, recall and F1 scores. The micro recall score indicates that for a random procedure, on average, we are 96.3% likely to correctly predict it. The micro precision score means that for a prediction, if we don’t consider the predicted class, there is a 93.5% chance that the predicted procedure code is correct. Furthermore, a high F1 score indicates that our model is working well overall. The micro scores of the model where slightly higher than the macro score, indicating that rare classes were being discriminated against. However, this representation and model achieved the highest score in all our metrics against every other representation and model.

Generalization check

We evaluated our best performing method’s ability to generalize to unseen sites, shown in Figure 2. When tested on a small number of sites, we observe low accuracy scores, with high variance. But as we increase the number of sites, the accuracy increases, and levels off at 80%, with ranges from 60% to 90%. This low initial score and high variance shows that generalization across sites is a valid concern when developing procedure code mappings. However, when we train our model on 17 or more sites, we observe that the our accuracy starts to level off at approximately 80%. The accuracy doesn’t reach the previously measured 96.3% as we were training on
a subset of the data that contained the originating site information. However, this evaluation shows that our model is able to generalize fairly well to completely new sites, which use unseen naming styles. However, we recognize that this dataset is obtained from a single geographical region, and between regions, naming standards can have higher variance. Thus, further work is needed in establishing the generalizability of this model on data from new regions.

Class Evaluation The confusion matrix, in Figure 3 shows a high-level view of the model’s performance. The thin yellow line across the diagonal shows that the model was strong at predicting classes. Since the confusion matrix was sorted with most common classes at the top left, and least common in the bottom right, we would expect to see the line fade as we move towards the bottom. Although fading is observable close to the bottom of the graph, the majority of the 1,855 classes are being correctly predicted.

Another error shown in this graph is the scattered yellow dots observable in the bottom left of the graph. These show that the model would occasionally predict the incorrect classes for some classes. The position of these scattered dots in the graph indicates that these classes were uncommon in the dataset and our model would predict common classes. Although, overall this graph shows that our model correctly identifies the majority of classes, it is evident that we were unable to correctly identify the correct class when given uncommon procedures.

Aggregate Class Evaluation Figure 4 shows two confusion matrices which evaluate whether the model is able to predict classes with the same modality or body part as the input. Both these confusion matrices have a strong diagonal indicating that the model is correctly able to identify the modality and body part of the input. The modality confusion matrix may have interesting behaviour between modalities, however, the low error rates makes it difficult to differentiate behaviour with random chance. However, the body part confusion matrix has higher error rates and presents more interesting behaviours. The model often confuses procedures relating to the head with procedures relating to the neck, or procedures capturing both head and neck. We also observe the same mistake between pelvis, abdomen and pelvis/abdomen procedures, however, to a lesser extent. Lastly, we observe high error rates with procedures with unspecified body parts. These errors in the model are understandable as these body parts are close to one another, or often combined into a single body part. Our errors in both precision and recall of unspecified body indicate that our model isn’t discriminating against or overcompensating predictions of unspecified body part, but rather makes mistakes by not recognizing the class, and mistaking input as that class. These confusion matrices show that even when our model makes mistakes, the prediction is semantically similar to the input, as the modality and body part of the procedure tend to be correct.
Figure 4: The figure on the left shows a confusion matrix of the model, when only observing the modality of the input and output, and the figure on the right shows the confusion matrix for body parts.

Qualitative evaluation of unsupervised clustering We found that the $k$-means grouped together semantically similar words. Cluster A and B grouped words related to extremity procedures and obstetrical ultrasound, respectively. Cluster C grouped together terms related to fluoroscopy and fluoroscopically-guided procedures - including variations in spelling (“fluoro” v “fluro”). While not objectively evaluated, subjective expert assessment suggests that we were able to learn clusters that were semantically meaningful. Future explorations is needed to more thoroughly evaluate the unsupervised clustering potential.

Table 3: Top 10 keywords for three example clusters. Each row represents a cluster.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Top keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>wrist, hand, elbow, cr, portable, shoulder, or, finger, right, uex</td>
</tr>
<tr>
<td>B</td>
<td>ultrasound, portable, single, abdpel, singleton, twins, triplets, obstetrical, us, dating</td>
</tr>
<tr>
<td>C</td>
<td>fluo, fluoro, fluoroscopic, fluoroscopy, fluro, flush, fna, fluid, embolization, rf</td>
</tr>
</tbody>
</table>

Impact

Semantic interoperability is one of the holy grails of medical informatics. Currently, cross site image sharing is dependent on laborious procedure mapping, which can be neglected. Depending on whether a site is joining a DIR, changing informatics systems, or introducing a new set of procedures, these manual mapping batches can range for a handful to hundreds of procedures. By building a model with 96.3% accuracy, we can enable high accuracy low effort mapping of procedure codes. This eliminates the hurdle surrounding manually mapping procedure codes, and allows sites to freely join a DIR, restructure their information system, and introduce new procedures and avoid the potential for gaps in care.

In practice, this model could be deployed under the supervision of an informaticist. As an example, the neural network model is able to provide the top $x$ predictions, as well as confidence for those predictions. This can enable a human augmented workflow, where high confidence procedures are automatically mapped, while low confidence procedures can be presented to a human with top $x$ suggested classes. Furthermore, these uncertain cases can be used to continue training the model, and increase robustness in the future.

Limitations and Future Work

Our procedure name mapping models were developed from a diverse, multi-site dataset; however, these were all located in one geographic region (Ontario, Canada), shared one language (English) and represent procedures reimbursed by a single payor. Generalizability of this model on other datasets would need to be tested, although we believe the approach is generalizable and transfer learning may facilitate.
In this paper, we described a solution to the problem of mapping local procedure codes to a regional terminology. This is most commonly a many to one mapping problem. The reverse problem - import and display of external prior (IDEP) images from a regional repository to a local site - may be even more important to imaging facilities. This could be solved by creating regional-to-local ML models at each site; transfer learning may help overcome the small local dataset problem. However, the one-to-many nature of IDEP would result in information loss.

We also believe that a model could be developed to support the deployment of a national or international standard terminology by using one standard (ie SNOWMED CT, Radlex-LOINC) as the labels for this supervised learning task.

Conclusion

Inaccurate and inefficient medical imaging exchange between imaging facilities is a pervasive problem in healthcare. In this paper, we present a real-world and novel application of NLP and ML to automate procedure name mapping using an Ontario-wide dataset. We explore several representations and models and measure their performance on our dataset. The best performing method we developed achieved 96.3% accuracy across 1,855 classes and demonstrated the ability to generalize to new data in 20 distinct sites. We also conduct an in-depth analysis of our best performing method, and show that the model correctly predicts most classes, despite the class imbalance in the dataset. We also show that when our model makes errors, the modality and body part tend to be accurate. This model enables effective image exchange, and opens up avenues for a widespread healthcare sharing system. We believe this approach can be applied to many interoperability challenges, including health information exchange, imaging facility consolidation and mapping medical lab tests or medical reports to a standard naming convention.

References


