Clustering Large-scale Diverse Electronic Medical Records to Aid Annotation for Generic Named Entity Recognition

Nithin Haridas
Carnegie Mellon University
Pittsburgh, Pennsylvania ‘
nithinh@cs.cmu.edu

Yubin Kim
UPMC Enterprises
Pittsburgh, Pennsylvania
kimy10@upmc.edu

ABSTRACT

The full extent of diversity in clinical documents and the effects on natural language processing (NLP) tasks in the medical domain have not been well studied. In supervised NLP tasks, it is vital to have training data that resembles test data [27]. In the medical domain, this often translates to uniform subject matter distribution [16]. We have access to a corpus of 157 million documents from 42 different electronic medical record (EMR) vendors, with over 40,000 distinct categories assigned to the documents. The sheer diversity of the documents is an obstacle to an accurate sub-sampling of the data for annotation. We propose that clustering clinical text documents is an effective way to aid the annotation effort and ensure coverage. We demonstrate the effect of lack of coverage in training data for a supervised generic named entity recognition(GNER) task and the impact of clustering on the task. We will also examine the characteristics of clusters generated from a diverse dataset.

KEYWORDS

clustering, generic named entity recognition, electronic medical records, diversity

This work was presented at the first Health Search and Data Mining Workshop (HSDM 2020)[7]

1 INTRODUCTION

In 2010, the American Recovery and Reinvestment Act was passed into law, requiring all public and private US healthcare providers to adopt electronic medical records (EMR) by January 1, 2014, empowering clinical information extraction and NLP research.

However, access to annotated data with a comprehensive coverage of subject matter domains remains a major challenge in clinical NLP. Widely available clinical document datasets are often small or are a narrow slice of the extant types of documents found in EMR systems. For example, the MIMIC dataset consists only of intensive care unit documents [14]. The i2b2/UTHealth 2014 dataset [24] is composed primarily of progress notes and discharge summaries. The dataset consists of 3 categories of patients at various stages of coronary artery disease.

Real world medical records are very diverse with respect to subject matter content and context [13]. There are lab procedures, consult notes, x-ray and ultrasound reports in cardiology, pulmonology, orthopedics to name a few.

We studied the data repository of a large healthcare provider from 42 different electronic medical record (EMR) vendors containing in excess of 157 million clinical text documents. EMR vendors are also referred to as **source systems**. The naming convention for categories assigned to the documents by the source systems are not necessarily consistent. We refer to this assigned category as a **document type**. The repository contains over 40,000 unique document types.

If we were to sample a set of documents across all the document types, we will end up with 400,000 documents with just 10 documents from each document type. In addition, annotations and their verification are done by subject matter experts. Evidently, this exercise requires large resources financially as well as with respect to time.

Generic named entity recognition is used to generate structured representation of a clinical text document by identifying biomedical concepts in the text. GNER can be used to extract hidden information in a diagnosis [5]. Information processing systems that rely on structured data cannot access such hidden information in clinical texts. The distribution of the biomedical concepts is dictated by the subject matter content of the document [6].

Supervised machine learning based GNER systems requires annotated clinical text documents with wide coverage [21] [4] [3]. Coverage in our context simply means that the training data contains a very diverse set of named entities. As we see above, this is a very challenging job when there are 40,000 categories. We propose that clustering the documents can delineate them into a smaller number of groups with each group aligned to similar subject matter content.

We examine the effects of inadequate coverage of training data for a generic named entity recognition (GNER) task and how clustering can mitigate some of these effects. Note that our objective is not to classify clinical text into a certain category, but to ensure coverage for the GNER task.

In this work, we want to answer the following research questions.

(1) How does lack of coverage in training data affect a supervised GNER system ?
(2) Can we cluster documents such that sampling from every cluster improves coverage ?
(3) How do we cluster documents and what are the characteristics of each cluster ?

In the following sections, we will describe the dataset in further detail, explain the clustering method that we used, experiments demonstrating the impact of lack of coverage in the GNER task and finally the impact of clustering on coverage.

2 RELATED WORK

Classification and clustering of electronic medical documents are relevant in other contexts within the medical domain. They are

primarily employed to address problems emanating from the diversity of documents based on subject matter content. BioASQ[1] orgainizes a large scale biomedical semantic indexing task every year to classify PubMed[2] documents into classes from the MeSH[3] hierarchy. Weng et al. [28] use a neural network architecture and a linear SVM for document classification in MGH [17] and iDASH [18] datasets respectively. The MGH dataset includes 3 subdomains (neurology, cardiology, endocrinology). iDASH is annotated with 6 subdomains (cardiology, endocrinology, nephrology, neurology, psychiatry and pulmonary disease).

Clustering has been used to extract medication and symptom names by Ling et al. [15] on the corpus from the i2b2/UTHealth 2014 dataset. Clustering has also been used for a drug repositioning task on a composite dataset consisting of 417 drugs and their properties by Hameed et al. [9]. The drug repositioning task identifies additional uses for certain drugs based on similarities with other drugs in the same cluster.

Features used in classification and clustering tasks in the medical domain frequently includes relevant terms extracted with the help of a concept mapper. cTakes, [20], MetaMap [1] and NobleCoder [26] are examples of concept mappers. A concept mapper uses predefined rules to identify medically relevant terms and retrieve a standard representation such as in UMLS [2]. Our work uses NobleCoder for this purpose. Section headers in a clinical document such as Complaint, Allergy and Summary have also been used as features in [15] and [8]. We also use section headers. We will see more details of possible set of features in Section 4.

Tang et al. [25] use clustering based word representation (WR) as features to improve a CRF based model in a biomedical named entity recognition (BNER) task on the BioCreAtIvE II GM [22] and JNLPBA [11] datasets. The BNER task is identical to the GNER task described in Section 5.2.

We employ clustering as a means to ensure diversity in annotated data in downstream tasks. As a result of the large number of document types from many source systems, it is not feasible to sample from every document type. The Document type dataset as we will see in Section 3.1 has 168 document types. We aim to show that sampling from clusters is a feasible strategy to represent diverse clinical text documents in annotated data. We specifically choose GNER as the downstream task because biomedical concepts are correlated with the relevant subject matter domain [6].

3 DATA

We describe here the corpus of the large healthcare provider to illustrate the scale of the problem. The data corpus is collected with ethical approval. The data processing pipeline anonymizes patient data. The processed data is stored in a HIPAA compliant environment with restricted access. We mentioned earlier that documents are assigned document types by source systems.

Naming document types depends upon 5 axes:

Subject matter domain: e.g. cardiology

Type of service: e.g. consultation

Kind of document: e.g. note, consent

Setting: e.g. hospital, clinic

Role: e.g. attending, consultant

However, the naming conventions are not uniformly applied even within the same system and source systems might not use all the axes when deciding to assign document types. Because of these incompatible naming conventions, our repository has a very large variety of document types.

The repository has 41,521 document types containing 718,337 documents. The data is diverse and the distribution across subject matter categories is not uniform. There are 32,468 types in IMAGECAST, for example, a source system primarily for radiology. This is because radiology documents have a different document type based on the relevant body part and the type of the image (X-ray, MRI).

4,458 of the document types had at least 100 documents (common document types) and the remaining types had less than 100 documents per type (sparse document types).

Among the 4,458 common document types, radiology notes were grouped based on certain conventions (such as first 2 letters of their name) into 18 types. The resulting consolidated dataset has 1296 common document types. We used this dataset to examine high level characteristics.

The most frequent tokens in the documents within a document type can give a certain idea about the document type. This is illustrated in Table 1.

We were able to identify certain patterns among document types and on closer inspection, we found that certain document types were duplicates of each other. The key properties were small euclidean distance between them in the feature vector space (described in Section 4, an high overlap of top terms within the document types' vocabulary and similar names for the document types. Some examples are shown in Table 2. Each duplicate pair here belong to the same source system. However, we cannot rule out a scenario where there are duplicate types across source systems. Clustering the documents can be a way to ensure that these duplicate note types are always grouped together which can significantly reduce annotation efforts.

The "GNER dataset" and the "Document type dataset" are subsets of the documents from the 1296 common document types.

3.1 Document type dataset

The Document type dataset is used to find the methodology for clustering and understand the best features. The resultant parameters are then used to cluster documents for the GNER dataset.

The dataset contains 13,440 documents spanning across 30 unique subject matter domains. The subject categories are listed in Appendix A. The Document type dataset's diversity is a very notable aspect. With 13,440 documents, it is smaller than the MIMIC [14] dataset(53,423 documents), but certainly more diverse in terms of subject matter content. To our understanding, we do not know of any other work that utilises clustering to tackle diversity in clinical text documents with broad coverage of subject matter content.

Subject matter content for documents in this dataset is informed by the document's document type. We can map many of the document types to standard representation from the subject matter domain axis of LOINC Ontology [13]. This mapping is first created

[1] http://bioasq.org/
[2] https://www.ncbi.nlm.nih.gov/pubmed
[3] https://meshb.nlm.nih.gov/search

id	Term 1	Term 2	Term 3	Category
1907	foot	ankle	incision	Orthopaedic Surgery
4906	liver	hepatitis	cirrhosis	Gastroentrology
106	artery	femoral	catheter	Cardiology

Table 1: Sample top terms in document types

Type 1	Type 2	distance	Overlapping top terms
BIDEXASKELNOREM	BIDEXASKEL	0.153	bone, density, mass
Neonatal_History	CDIDNUM	0.09	infant, birth, delivery
IM_Office_Visit	FP_Office_Visit	0.192	continued, take, encounter
ED Note	HP_ED_Note	0.21	active, scope, coding

Table 2: Duplicate note types in the dataset.

id	AP	CE	CO	CV	EP	IM	MU	PR	VA
0	0	6	21	0	0	496	0	0	0
1	0	0	0	0	98	0	0	0	0
2	0	0	183	0	0	0	0	0	0
3	35	529	21	12	2	36	21	7	0
4	0	0	0	0	0	0	0	172	0
5	0	269	0	0	0	0	0	0	0
6	4	3	0	85	0	6	0	0	0

Table 3: Distribution of source systems in the clusters for the GNER dataset AP: APOLLO, CE: CERNER, CO:COPATH, CV:CVIS, EP:EPIC, IM:IMAGECAST, MU: MUSE, PR: PROVATION, VA:VASCUPRO

by a data analyst and subsequently verified by a subject matter expert. The 13,440 documents belong to 168 document types across the 30 subject categories.

3.2 GNER dataset

Our experiments for the GNER task uses a dataset of 2059 documents. We refer to this as the **"GNER dataset"**. The GNER dataset consists of documents with each token annotated as one of G-B, G-I or O. All generic mentions of medical concepts (e.g. conditions, procedures, labs-observation, medications) are annotated.

We do not have information about the subject categories for these documents. We do however, know the source systems for the documents. The distribution of source systems in the dataset is shown in table 3. Documents from different source systems have notable differences in content. COPATH notes are clinical observations and procedures while IMAGECAST notes are radiology observations and procedures. CERNER and EPIC are typically systems used at hospital level and contains a mixture of note types such as progress notes, consults and discharge summaries. PROVATION notes are gastroenterology procedures while VASCUPRO notes are vascular lab reports. MUSE notes consists of cardiology procedures, CVIS notes are related to cardiac imaging and APOLLO notes are for documenting physical therapy.

4 CLUSTER GENERATION PIPELINE

The motivation behind clustering is that it can create coherent groups of data. We show how closely the clusters are aligned according to their subject matter domain. For this analysis, we use the "Document types dataset".

4.1 Purity

Clusters are evaluated based on how coherent they are to a particular subject matter. This is measured using the purity metric. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N, the total number of documents.

$$purity(W, C) = \frac{1}{N} \sum_{k} \max_{j} |w_k \cap c_j|$$

where $W = (w_1, w_2, w_3, ...w_k)$ is the set of clusters and $C = (c_1, c_2, c_3, ..c_j)$ is the set of classes.

Let us look at the specific steps involved in generating document clusters.

4.2 Preprocessing and tokenizing documents

All documents are preprocessed for stopword removal (stopwords from the nltk[4] corpus), filtering out highest frequency words (top

[4]https://www.nltk.org/

0.1%) in the corpus and words with only numbers in them. The documents are then tokenized with nltk word tokenizer. The number of unique tokens in the dataset after preprocessing and tokenizing is approximately 2.7 million.

4.3 Feature generation

4.3.1 N-gram word tokens. The word tokens generated in the previous process were used to generate unigram, bigram and trigram word tokens. Looking at the top terms for documents, they indicate the subject matter domain (e.g. cardiology, neurology) or the kind of document (e.g. note, letter) in most of the cases. Bigrams and trigrams are included as features because many medical concepts are 2 or 3 words or more. Some examples are “intravenous solution" and “atrial sinus solitus". We only considered the top 200,000 unigrams based on document frequency. As for bigrams and trigrams, the total size was 8.8 million and 23.5 million respectively. To limit the size, we put additional restrictions based on corpus frequency (greater than 30).

4.3.2 N-gram word tokens filtered on medical vocabulary. We also restricted the word tokens to only those found in a medical dictionary[5]. The medical terms are unigram tokens from two corpuses (OpenMedSpel[6] and MTH-Med-Spel-Chek[7]). Bigrams and trigrams are selected only when all the constituent terms are part of the dictionary. This was shown to group documents of the same subject categories even closer. The clustering purity based on unigrams with no filtering on the “Document type dataset" was 0.56 while the one based on filtered unigrams was 0.63. It also had the considerable advantage of reducing the dimensionality of the features. Unigram frequency reduced from 200,000 to 17,462. Bigram and trigram frequencies were 100,000 and 70,000 respectively (Table 4).

4.3.3 Section headers. Medical records typically have sections that distinguish one type from another. For e.g. A cytology report contains section headers such as “CLINICAL HISTORY" and “SOURCE OF SPECIMEN", whereas a test report of lumbar puncture has “PROCEDURE" and “TECHNIQUE". These section headers act as a signature for documents from the same source and can be used to identify format level features for the documents. We extract section headers from the documents with NobleCoder [26] and tokenize them with nltk word tokenizer. Tokenizing the section headers helps create unigram, bigram and trigram features from the section headers that are normalized across the documents.

4.3.4 Concepts. NobleCoder tool [26] is able to map words pertaining to medical concepts into their UMLS [2] representation. In addition, we can restrict the concepts to be only from certain semantic types. With inputs from knowledge engineers, we only extracted concepts from a particular list of semantic types. For details, refer Appendix B. Similar to the use of medical vocabulary, we can reduce our focus to medical concepts only when identifying features.

[5] github.com/glutanimate/wordlist-medicalterms-en
[6] https://e-medtools.com/
[7] https://rajn.co/

Feature	Count
Unigrams	17462
Bigrams	100000
Trigrams	70000
Section headers	8432
Concept tokens	27453

Table 4: Feature count for multiple categories

4.4 Tf-idf matrix for generated features

N-gram tokens, section headers and concepts are combined and they are weighted by their tf-idf (Term frequency - Inverse document frequency) [23] values. 3 separate feature matrices are created corresponding to n-gram features, section headers and concepts respectively.

The feature matrices are then horizontally stacked[8] in a variety of combinations to examine their effects (5.1) Horizontal stacking is simply concatenating the individual feature matrices row-wise. E.g. a 13,440 x 8,432 feature matrix for section headers is combined with a 13,440 x 27,453 concept feature matrix to get a 13,440 x 35,975 combined feature matrix.

4.5 K-means clustering

In K-means clustering, (Hartigan and Wong [10]) datapoints are divided into clusters of equal variance. K-means clustering is a fast algorithm and the speed allows us to iterate quickly based on different combinations of features. Document similarity is determined based on the euclidean distance between the corresponding feature vectors.

K-means clustering initializes ‘k’ centroids when generating ‘k’ clusters. Each centroid will be a particular document in the dataset. In our setting, the centroid is chosen at random for each execution of the algorithm. Predetermined centroids did not change the purity values for the generated clusters. Subsequent members for each cluster is chosen when that document is closer to a particular centroid than other centroids. The centroids are then recalculated after adding each member. The algorithm is run multiple times and we check whether the results are consistent across each run. We tried a wide range of k-values from 10 to 140, but the clusters were found to be fairly stable and consistent across all values. We also tried other clustering techniques and the resultant purity values were comparable to values obtained from the k-means method.

Table 5 illustrates the content for some of these clusters. Cluster 0 is about sleep medicine and cluster 2 is about x-rays. But clusters 3 and 4 are about consults and office visits.

5 EXPERIMENTS

5.1 Examining the best features for clustering for subject matter domain

Each document in the Document types dataset corresponds to a sample datapoint for the clustering algorithm. The Document type dataset has 13,440 samples. Each document has a subject matter category assignment and there are 30 unique subject categories. The

[8] https://docs.scipy.org/doc/scipy/reference/generated /scipy.sparse.hstack.html

id	Term 1	Term 2	Term 3
0	apnea	events	rem
1	ms	ear	density
2	xray-left	vendor	x-ray right
3	reference	discharging	client
4	recorded	routing	best practice

Table 5: Sample top terms in clusters

Features used	Purity
Random	0.16
Source	0.44
Unigrams	0.63
Section headers	0.53
Concept tokens	0.57
Unigrams and bigrams	**0.64**
Unigrams, bigrams, trigrams	0.61
Unigrams and section headers	0.60
Section headers and concepts	0.62
All features	0.59

Table 6: Purity comparison between clusters generated from particular features

number of clusters are chosen as 30 based on the mean silhouette coefficient of all samples. Silhouette coefficient of a sample i in a cluster C is defined as

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}, |C_i| > 1$$

$$s(i) = 0, |C_i| = 1$$

where a_i is the mean distance of i to other points in C and b_i is the smallest mean distance of i to all points in any other cluster, of which i is not a member.

The “closeness" of samples in a cluster can be measured by a cluster's mean silhouette coefficient. High silhouette coefficient of a sample in a cluster [19] indicates that it is a part of a cluster with similar samples and separated from dissimilar samples.

Thus, we use the following parameters for the k-means algorithm, number of runs:10; number of clusters: 30. The “default" information we have about the document types is their source systems. So our baseline purity based on clustering by the 24 source systems in the dataset is 0.44 (Table 6). We use all the features we mentioned in previous sections, namely, unigrams, bigrams, trigrams, section headers, and concept tokens. Each of these features independently outperform the baseline. On the basis of the experiments, we see that simply using unigrams and bigrams (117,462 features) resulted in the highest values. Using unigrams (17,462) on its own or the combination of section headers and concepts (35,975) have comparable purity. Usage of the latter set of features have an added advantage of lower dimensionality as the corresponding feature vector is much smaller. For the experiments on the GNER dataset, we only used unigram features.

Taken out	Dev F1	Test F1	Size
APOLLO	44.7	8.1	39
CERNERH1	52.1	13.6	807
COPATH	33.9	4.9	225
CVIS	43	27.4	97
EPIC	46.4	24	98
IMAGECAST	36.9	5.5	538
MUSE	44.6	0	21
PROVATION	37.3	21.2	179
VASCUPRO	48.7	0	51

Table 7: Effects of source systems on the GNER task. F1 score in percentage. Size refers to size of the test set (taken out source).

5.2 Generic Named Entity Recognition Task

The GNER task is modeled as a sequence labeling problem. Given a word from the document, the task is to predict one of the labels. The labels used are G-B = beginning of an entity, G-I = inside an entity, and O = outside of an entity. For e.g., consider the input sentence, "The patient is suffering from a cardio-vascular disease." The correct predictions would be (The,O), (patient,O), (is, O), (suffering,O), (from,O), (a,O), (cardio,G-B), (vascular,G-I), (disease,G-I),(.O).

GNER is an important step in the NLP pipeline to extract biomedical entities and concepts. The GNER model we use is a Bi-LSTM-CRF based model (Huang et al. [12]). The Bi-LSTM layer has access to past and future tokens at any particular time-step in the token stream. The CRF layer captures sentence level information. We make no changes to the configuration and features for the Bi-LSTM-CRF model from the version presented in [12].

We use the GNER dataset for this task. Recall that we do not have access to subject matter annotations for documents in the GNER dataset. But we find that, in this dataset, the document's source system is a good indicator of subject matter content (Section 3.2). The lack of coverage in training data is first illustrated using document's source system as a proxy for subject matter content. We will then show that if we use a document's cluster in place of the source system, GNER system performance suffers in a similar manner. We also use the clustering methodology as explained in Section 4. Performance of the GNER model is measured as the F1 score on the test data.

There are 3 parts to the experiments in the GNER task.

(1) **Leave one out cross validation based on document's source system**. This identifies the drop in performance for documents in test data that are from unseen source systems in training.
(2) Cluster the documents. **Leave one out cross validation based on document's cluster**. We show that performance of GNER on documents in test data that are from unseen clusters in training similarly drops
(3) **Train and test the documents on a per cluster basis**. Finally, we examine the GNER performance on documents trained and tested on a per cluster basis.

5.2.1 Effect of source systems on the GNER task. Documents in different source systems belong to different sub-domains and their

format is different as well. We investigate the effect of source systems on GNER performance with leave one out cross validation on the dataset, based on the source system of the documents. Documents originating from one source system is left out and kept aside as a test set. Remaining documents are shuffled to make a train/dev split in the ratio 9:1. The trained models are tested on the corresponding dev set and the corresponding left out test set.

Table 7 shows the effect of source systems on GNER performance. Performance of all 8 trained models on the dev set is much better than their performance on the left out test set. Documents from source systems unseen in training causes a drop in performance. We will refer to this as "the unseen type problem". This indicates that when creating training data, we want to have as broad a coverage as possible. It is not feasible to annotate documents across the overwhelming amount of document types that are available. We will see how clustering the documents ensures that the document types being selected are sufficiently diverse.

5.2.2 Effect of clustering on GNER task. For the next part of our experiment, we created 7 clusters from documents in the GNER dataset. Table 3 shows the distribution of the document's source systems in these 7 clusters. Each of the seven clusters are dominated by documents from a particular system. With the exception of cluster 3 and the source system MUSE, we can almost map a source system with a particular cluster. The features and methodology for clustering is described in detail in Section 4. We perform leave one out cross validation on the 7 clusters. One cluster is selected, and kept aside as the test set. The remaining documents in rest of the documents is split into train/dev sets in the ratio 9:1. Then we train the GNER model on the resulting training set and test the model on the dev set and test sets. This is repeated for every cluster.

We see that there is a drop off in test set performance with the exception of cluster 2 (Table 9). The test performance is low for clusters that are strongly dominated by one source with no or very little documents of that source in other clusters. These are clusters 1,4, and 6. For cluster 0, the source systems of documents in this cluster, CERNER, COPATH and, IMAGECAST are also represented in other clusters. In the case of clusters 3 and 5, the performance is lower (but still higher than 4 other clusters) because it is dominated by documents from CERNER. Even though there are 278 documents from CERNER in the training set, there is a diverse range of documents from that system.

For cluster 2, which is dominated by COPATH documents, there are still 42 documents in training data. Further more, the documents in COPATH are mostly clinical pathology or observation notes with a limited vocabulary. They also tend to be very dense with many generic named entity examples to learn from. For instance, when training on a per cluster basis, cluster 2 has 3857 named entities in the training set from 200 documents. There are only 761 entities from 460 documents in the training set for cluster 0.

Going back to the 7 clusters in the leave one out cross validation, we generated train/test split for each of the clusters. The GNER model was trained and tested on a per cluster basis. Table 10 shows the results for the GNER model when it is trained and tested on documents from each cluster separately. The F1 scores in the test set are better in 3 of the 7 cases, maintained for 2 cases and drops for clusters 1 and 6.

size	silhouette score avg	Test F1 (%)
181	-0.0059	27.99
431	-0.071	30
175	0.082	35.22
796	0.143	70.23
478	0.209	55.64

Table 8: Avg silhouette score of samples in a cluster vs Test F1 score. Pearson coefficient 0.63

Taken out	Dev F1(%)	Test F1(%)
0	46.7	36
1	43.8	18.4
2	35.8	54.6
3	44.4	23.1
4	43.4	14.7
5	40.4	19
6	45.5	15.1

Table 9: Effects of clusters on the GNER task

Cluster id	Train F1(%)	Test F1(%)
0	44.60	36.36
1	66.09	11.86
2	71.36	78.70
3	64.23	23.30
4	69.29	55.93
5	66.49	32.41
6	0	0

Table 10: GNER model trained and tested on a per cluster basis

Cluster 2 has a high performance because of the nature of the clinical pathology documents that constitute the cluster. Cluster 4 is dominated by documents from PROVATION which also shares some of the properties of cluster 2. Both clusters have a limited vocabulary of biomedical concepts with a dense distribution within the documents. Cluster 6 has a training and test F1 of 0 because there are only 14 named entities in the entire training set. This brings to a potential pitfall when training on a per cluster basis. We also need to make sure that each cluster has enough training data. Cluster 1 on the other hand, has enough training data, but it is dominated by documents from the EPIC source system, which similar to CERNER is very diverse in terms of subject matter domain. This is the second possible pitfall. We need to choose the features so that clusters are aligned on the subject matter domain as much as possible. CERNER and EPIC documents are grouped together despite the diverse nature of constituent documents because general hospital notes are closer to each other than say, radiology notes.

We looked at 5 clusters generated from the GNER dataset, divided each cluster into train and test sets. We trained the Bi-LSTM-CRF model on the training sets of each cluster separately and tested it on the corresponding test sets from that cluster. We posit that the

silhouette coefficient of a cluster is correlated with the GNER performance. In Table 8, the average silhouette coefficients of samples in a cluster are compared against the GNER F1 score when trained and tested on a train/test split from the same cluster.

The high pearson correlation coefficient of 0.63 indicates that well formed clusters have a high linear correlation with a higher test GNER performance. We saw in Section 4 that the "closeness" of the samples in a cluster is influenced by their respective subject matter domain.

Despite the caveats mentioned above, when we choose documents for annotation, we can improve test performance by choosing them from as many clusters as possible. This avoids having to annotate subject matter domain for each document type.

6 CONCLUSION

The diversity of unstructured clinical text documents has been an under-studied problem in clinical NLP. This paper presented initial explorations into a large real-life data repository of 157 million documents across 42 source systems and found that the source systems reported more than 40,000 document types.

Initial explorations of the document types showed that they vary widely in content and format, with significant ramifications on supervised NLP tasks. When a supervised generic named entity detection model was tested on document types that had not been present in the training data, the performance is much lower compared to a model trained on a more diverse training set ("the unseen type problem"). This indicates a need for careful selection of data when annotating to create a training set for a new NLP task in a real world setting; poorly chosen training data will hinder the creation of generalizable models. Ideally, an annotated training set would have coverage over all subject matter content. However, due to the large number of document types available, this is prohibitively expensive. Our study showed that many of the types reported by the systems are actually quite similar, leading us to explore clustering as a method to mitigate diversity of note types.

We experimented with various features for clustering and found that to generate clusters along subject matter domains, a combination of unigram and bigram features worked well, providing high purity scores on the "Document types dataset". Since we do not have subject matter annotations for the larger data repository of 40,000 document types, we posit that clusters are a reasonable stand-in to ensure representation.

We showed that clustering captures information that helps translate training performance for the GNER task. By clustering the document types to a smaller set of clusters, it is possible to select training data for NLP tasks with good coverage without wasting annotation effort on similar types.

Future work may include utilizing semantic embedding representations of documents and training feature weights for better clustering.

REFERENCES

[1] Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

[2] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

[3] Andreea Bodnari, Louise Deléger, Thomas Lavergne, Aurélie Névéol, and Pierre Zweigenbaum. 2013. A supervised named-entity extraction system for medical text. In *CLEF*.

[4] K. P. Chodey and G. Hu. 2016. Clinical text analysis using machine learning methods. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–6.

[5] Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760 – 772. Biomedical Natural Language Processing.

[6] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

[7] Carsten Eickhoff, Yubin Kim, and Ryen White. 2020. Overview of the health search and data mining (hsdm 2020) workshop. In *Proceedings of the Thirteenth ACM International Conference on Web Search and Data Mining*, WSDM '20, New York, NY, USA. ACM.

[8] K. Ganesan and M. Subotin. 2014. A general supervised approach to segmentation of clinical texts. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 33–40.

[9] Pathima Nusrath Hameed, Karin Verspoor, Snezana Kusljic, and Saman Halgamuge. 2018. A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration. *BMC Bioinformatics*, 19(1):129.

[10] J. A. Hartigan and M. A. Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

[11] Ming-Siang Huang, Po-Ting Lai, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2019. Revised JNLPBA corpus: A revised version of biomedical NER corpus for relation extraction task. *CoRR*, abs/1901.10219.

[12] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

[13] Stanley M. Huff, Roberto A. Rocha, Clement J. McDonald, Georges J. E. De Moor, Tom Fiers, Jr. Bidgood, W. Dean, Arden W. Forrey, William G. Francis, Wayne R. Tracy, Dennis Leavelle, Frank Stalling, Brian Griffin, Pat Maloney, Diane Leland, Linda Charles, Kathy Hutchins, and John Baenziger. 1998. Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary. *Journal of the American Medical Informatics Association*, 5(3):276–292.

[14] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035 EP –.

[15] Y. Ling, X. Pan, G. Li*, and X. Hu. 2015. Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization. *IEEE Transactions on NanoBioscience*, 14(5):500–504.

[16] Georgia McGaughey, W Patrick Walters, and Brian Goldman. 2016. Understanding covariate shift in model performance. *F1000Research*, 5:Chem Inf Sci–597.

[17] Shawn N Murphy and Henry C Chueh. 2002. A security architecture for query tools used to access large biomedical databases. *Proceedings. AMIA Symposium*, pages 552–556.

[18] Lucila Ohno-Machado, Vineet Bafna, Aziz A Boxwala, Brian E Chapman, Wendy W Chapman, Kamalika Chaudhuri, Michele E Day, Claudiu Farcas, Nathaniel D Heintzman, Xiaoqian Jiang, Hyeoneui Kim, Jihoon Kim, Michael E Matheny, Frederic S Resnic, Staal A Vinterbo, , and the iDASH team. 2011. iDASH: integrating data for analysis, anonymization, and sharing. *Journal of the American Medical Informatics Association*, 19(2):196–201.

[19] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

[20] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin 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 : JAMIA*, 17 5:507–13.

[21] M. Shekhar, V. R. Chikka, L. Thomas, S. Mandhan, and K. Karlapalem. 2015. Identifying medical terms related to specific diseases. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 170–177.

[22] Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A Struble, Richard J Povinelli, Andreas Vlachos, Jr Baumgartner, William A, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W John Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9 Suppl 2(Suppl 2):S2–S2.

[23] Karen Sparck Jones. 1988. Document retrieval systems. In Peter Willett, editor, *Document Retrieval Systems*, chapter A Statistical Interpretation of Term Specificity and Its Application in Retrieval, pages 132–142. Taylor Graham Publishing, London, UK, UK.

[24] Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *Journal of biomedical informatics*, 58 Suppl(Suppl):S67–S77.
[25] Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. 2014. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014:240403.
[26] Eugene Tseytlin, Kevin Mitchell, Elizabeth Legowski, Julia Corrigan, Girish Chavan, and Rebecca S. Jacobson. 2016. NOBLE – flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*, 17(1).
[27] Grace. Wahba. 1990. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics.
[28] Wei-Hung Weng, Kavishwar B. Wagholikar, Alexa T. McCray, Peter Szolovits, and Henry C. Chueh. 2017. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making*, 17(1):155.

A GOLD ANNOTATED SUBJECT MATTER CATEGORIES

Gastroentrology
Surgery
Podiatry
Physical Medicine and Rehabilitation
Pulmonary Medicine
Orthopaedic surgery
Surgical Oncology
Cardiology
Family Medicine
Allergy
Molecular Genetic Pathology
Anesthesiology
Diagnostic Radiology
Otolaryngology
Neonatal perinatal summary
Interventional Radiology
Geriatric Medicine
Nuclear Medicine
Emergency Medicine
Neurology
Endocrinology
Obstetrics and Gynecology
Clinical Pathology
Sleep Medicine
Radiation Oncology
Hematology
Mental Health
Urology
Rheumatology

B SEMANTIC TYPES FOR CONCEPT FILTERING

Category
Health Care Related Organization
Gene or Genome
Congenital Abnormality
Acquired Abnormality
Clinical Drug
Body System
Cell Component
Body Location or Region
Injury or Poisoning
Body Space or Junction
Hazardous or Poisonous substance
Finding
Laboratory or Test Result
Pathologic Function
Cell
Virus
Therapeutic or Preventive Procedure
Fungus
Mental or Behavioral Dysfunction
Anatomical Abnormality
Bacterium
Neoplastic Process
Body Part, Organ, or Organ Component
Biomedical or Dental Material
Anatomical Structure
Disease or Syndrome
Indicator, Reagent, or Diagnostic Aid
Organic Chemical
Sign or Symptom
Occupation or Discipline
Pharmacologic Substance
Biomedical Occupation or Discipline
Diagnostic Procedure
Social Behavior
Laboratory Procedure
Tissue