Horses to Zebras: Ontology-Guided Data Augmentation and Synthesis for ICD-9 Coding

Horses to Zebras: Ontology-Guided Data Augmentation and Synthesis for ICD-9 Coding

Matúš Falis[1], Hang Dong[2], Alexandra Birch[1], and Beatrice Alex[3,4]

[1]Institute for Language, Cognition and Computation, University of Edinburgh
[2]Department of Computer Science, University of Oxford
[3]School of Literatures, Languages and Cultures, University of Edinburgh
[4]Edinburgh Futures Institute, University of Edinburgh
{s1206296, a.birch, balex}@ed.ac.uk, hang.dong@cs.ox.ac.uk

Abstract

Medical document coding is the process of assigning labels from a structured label space (ontology – *e.g.*, ICD-9) to medical documents. This process is laborious, costly, and error-prone. In recent years, efforts have been made to automate this process with neural models. The label spaces are large (in the order of thousands of labels) and follow a big-head long-tail label distribution, giving rise to few-shot and zero-shot scenarios. Previous efforts tried to address these scenarios within the model, leading to improvements on rare labels, but worse results on frequent ones. We propose data augmentation and synthesis techniques in order to address these scenarios. We further introduce an analysis technique for this setting inspired by confusion matrices. This analysis technique points to the positive impact of data augmentation and synthesis, but also highlights more general issues of confusion within families of codes, and underprediction.

1 Introduction

Large-Scale Multi-Labelled Text Classification (LMTC) tasks, such as automated ICD-9 coding of MIMIC-III discharge summaries, suffer from a big-head long-tail distribution of classes. This phenomenon naturally arises due to some labels being more frequent than others. This can further be affected by the source of the data – in the case of clinical *Natural Language Processing* (NLP), this is often a single institution. For instance, hospitals in Switzerland are unlikely to have cases of injuries caused by shark bites (code *W56.41XD* in ICD-10). Hence, depending on the data source, some labels will have a very small population – *Few-Shot* (FS), or even no population at all – *Zero-Shot* (ZS) scenario. Furthermore, adding new labels into a standard by splitting/fusing/altering existing concepts, or introducing new concepts also creates a ZS scenario. Medical coding methods need to be able to adapt to these scenarios.

Medical coding methods can be broadly divided on the task (document-level/entity-level) and the approaches (rule-based/machine learning) they use. Methods like Apache cTAKES (Savova et al., 2010), MedCAT (Kraljevic et al., 2019), or SemEHR (Wu et al., 2018) perform *Named Entity Recognition and Linking* (NER+L), identifying specific spans of text within the document and associating them with concepts in a knowledge base, such as UMLS (Bodenreider, 2004). They primarily use rule-based methods, with some inclusion of machine learning – *e.g.*, MedCAT uses contextual word embeddings for disambiguating homographic strings, such as context-sensitive abbreviations (*e.g.*, the string *"HR"* can mean "hour" or "heart rate"). Neural LMTC models, such as CAML (Mullenbach et al., 2018), predict labels on the document level. Labels are not associated with any particular string in the text, but rather appear as document-level sets.

Rule-based NER+L methods, assuming machine learning is not used, are not affected by the FS/ZS scenarios. There either is a suitable rule designed for a given situation, or not. If a new code is introduced into the label space, the rules need to be adjusted to reflect this.

Neural learning approaches are data-driven. The populations of labels available during training and the variety in the inputs to which they are associated affect the model's generalisability, especially if the model is not designed with the few-shot/zero-shot scenario in mind. Previous work has tried to address this with setting non-trainable parameters within the network as representations of ICD-9 codes enriched with knowledge from the ontologies (Rios and Kavuluru, 2018). While the few-shot/zero-shot performance improved, the overall performance deteriorated.

An alternative to model adjustments is to avoid the FS/ZS scenarios by supplying more data, *i.e.*, through data augmentation or synthesis. Aug-

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 389–401
May 26, 2022.

mentation through synonym replacement has been previously done using WordNet (Ollagnier and Williams, 2020), with improvements coming from the use of a medical knowledge base (UMLS) (Kang et al., 2021). Simple natural language generation techniques were also employed (Ollagnier and Williams, 2020). These techniques, while expanding the vocabulary, are only capable of producing synthetic documents with labels present in the original training data. Synthesising new documents with alternative labels has been done based on document templates in the scope of radiology reports – however, human experts were involved in the process (Schrempf et al., 2020).

We propose a novel type of data synthesis for ICD-9 coding, and medical LMTC tasks in general with the aim to replace concepts of underspecified codes with more specific, and often less frequent, alternatives. Similar to Schrempf et al. (2020) we recognise the value of augmenting concepts of interest. Rather than using templating in order to determine concept location, we use pre-existing NER+L techniques (MedCAT and SemEHR) to identify spans relevant to the gold standard labelling.

Furthermore, we introduce an error analysis technique for this setting inspired by confusion matrices. This technique associates codes within the prediction set with codes in the gold standard set according to the ontological structure allowing us to track mispredictions co-occurring with unmatched gold-standard codes indicating confusion – which codes tend to be mispredicted as others.

Our work provides the following contributions:

- Applying *Ontology-Guided Synonym Replacement* to ICD-9 coding, where multiple ontologies are used to determine candidate synonyms for a given concept found by an NER+L method akin to the work of (Kang et al., 2021). This *augmentation* method leads to improved model performance.
- *Sibling-Code Replacement*, where the surface form of a concept reported by an NER+L method is replaced with one of a semantically similar code according to the ontology, with the change being reflected in the document's updated silver standard. This *synthesis* method leads to improved model performance.
- The *Weak Hierarchical Confusion Matrix* (WHCM) – an analysis tool for the LMTC (weakly-labelled) scenario inspired by the concept of confusion matrix allowing more in-depth error analysis facilitating further development of LMTC systems. The output of this tool can be further used as an evaluation metric describing error types.
- The source code for augmentation and synthesis[1], and WHCM[2] will be made available via GitHub.

Our augmentation and synthesis methods both separately and combined lead to improved micro-F1 scores. The also improve g FS and ZS performance – although are surpassed by the baseline setup with more training. Our analysis tool highlights the error types in prediction – some errors are due to confusion within the code family, but most are due to underprediction.

2 Background

In this section we will introduce medical ontologies, both as a label set, and source of external knowledge. We will describe Named Entity Recognition and Linking used for determining relevant spans of text, introduce LMTC as our task, discuss previous data augmentation techniques in clinical NLP, and finally comment on the current approaches to evaluation and analysis of LMTC models.

2.1 Medical Ontologies

The International Classification of Diseases 9th Edition, Clinical Modification[3] (ICD-9-CM, here refereed to as ICD-9 despite nuances) is a medical ontology of diseases and procedures represented by two tree-structured label-spaces. The higher the depth of a node within the label space, the more specific a concept it describes, with lower depths representing aggregation on *e.g.*, disease type or general anatomy. Such aggregation is represented via subtrees (or *families*) of codes. Coding is done primarily with leaf nodes, representing the highest degree of specification within the ontology. We use ICD-9 as a basis for our research due to the availability of data labelled with this ontology – MIMIC-III. Newer revisions of the ICD (ICD-10[4], ICD-11[5]) differ in size, organisation of the tree

[1] https://github.com/modr00cka/Ontology-Guided-Augmentation-and-Synthesis

[2] https://github.com/modr00cka/weak_hierarchical_confusion

[3] https://www.cdc.gov/nchs/icd/icd9cm.htm

[4] https://icd.who.int/browse10/2019/en

[5] https://icd.who.int/browse11/l-m/en

structure, and naming conventions, but generally follow the same structural design principles. Hence our research can be re-used for newer standards.

An ICD-9 code (*e.g.*, *250.01*) consists of a *category* (part of the code appearing prior to the decimal point, *e.g.*, 250) and *etiology* (appearing after the decimal point, *e.g.*, 01). The etiology can be represented by up to two digits. A longer etiology implies a more specific concept.

Dedicated leaf-level codes exist to describe an "unspecified" version of a parent concept (*e.g.*, hypertension with unspecified malignancy status would be coded as *401.9 Unspecified Essential Hypertension*, rather than *401 Essential Hypertension*). Such "unspecified" concepts may appear on different depths representing different parts of the concept being unspecified. This phenomenon can appear within the same family of codes, indicating different levels of specificity – *e.g.*, the single-digit-etiology leaf code *365.9 Unspecified Glaucoma* versus the double-digit-etiology leaf code *365.60 Glaucoma associated with unspecified ocular disorder*. While not a general rule, some etiology patterns tend to be associated with unspecified concepts – *.9*, *.?0*, and sometimes *.?1* (where *?* can be any digit.)

The *Unified Medical Language System* (UMLS)[6] (Bodenreider, 2004) is a project of medical terminology originally released in 1990. The core components of UMLS are the *Metathesaurus* containing various medical vocabularies, a *Semantic Network* representing the connections between the terms, and an *Information Sources Map*. The concepts within the Metathesaurus are each assigned an identification code known as the *Concept Unique Identifier* (CUI).

Furthermore, the Information Sources Map component enables mapping of concepts between ontologies through the concepts' CUI. An examples of such a mapping is the SNOMED CT[7] to ICD-9 map curated by UMLS.

2.2 Named Entity Recognition and Linking

The task of identifying relevant concepts within free text is known as *Named Entity Recognition* (NER). It can be extended to NER+L by linking them to entities in an ontology (*e.g.*, UMLS). The standard labelling in NER+L tasks consist of two pieces of data – the indices identifying the span of text constituting an entity, and the assigned class (*e.g.*, CUI). NER+L serves as the first step in our augmentation and synthesis methods.

In the medical domain, notable early NER+L (predominantly rule-based) systems include MetaMap (Aronson, 2001) and Apache cTAKES (Savova et al., 2010). These systems struggle with ambiguities and spelling mistakes. BioYODIE (Gorrell et al., 2018), a more recent approach, addresses some of these ambiguity issues through corpus-based statistics, *e.g.*, co-occurrence graph. SemEHR (Wu et al., 2018) improves upon the output of BioYODIE with manually-derived rules. Certain types of ambiguity still pose issues to these systems, *e.g.*, expansion of context-sensitive abbreviations and variety of concept names. MedCAT (Kraljevic et al., 2019) employs unsupervised training and vocabulary building to further address ambiguity through context-sensitive disambiguation.

2.3 Large-Scale Multi-Label Text Classification (LMTC)

Large-Scale Multi-Label Text Classification (LMTC) is the task of assigning multiple weak labels to text documents. The labels come from a large label-space (in the order of thousands of labels), which can be structured, *e.g.*, ICD-9. LMTC tasks appear in several domains, including medical, legal, and commercial. The most notable early model in LMTC is CAML introduced by Mullenbach et al. (2018) for ICD-9 coding of medical documents.

Given an input document the model identifies the set of labels to be assigned. The input tokens are converted into word embeddings using word2vec (Mikolov et al., 2013). Convolutional filters are applied on these embeddings for short-range interaction. These phrase embeddings are fed into a label-specific attention mechanism – for each label an attention mechanism is applied identifying tokens contributing towards the respective code's prediction. The attention is multiplied with the the phrase embeddings resulting in label-specific document embeddings upon which classification is performed. Mullenbach et al. (2018) used spans around high-attention tokens (keywords) for qualitative evaluation of predictions. Falis et al. (2019) with the use of a hierarchical ensemble showed that tokens relevant for a family of codes can be captured with the attention mechanisms of ancestor

[6] https://www.nlm.nih.gov/research/umls/index.html

[7] https://www.snomed.org/

codes and propagated to descendants.

LMTC models are data-driven neural approaches requiring large amounts of data. Due to the big-head long-tail label space, the performance of these models varies between codes, with frequent codes performing better. For this reason FS and ZS specific techniques were developed, such as that of Rios and Kavuluru (2018); Lu et al. (2020).

2.4 Data Augmentation

Data Augmentation (DA) in machine learning is a method for artificially increasing the amount of training data by label-preserving alterations of the input. This technique can be used either to make the models more resilient to noise in the data, introduce variety, or enrich with additional information addressing model limitations.

One of the most representative DA techniques in NLP is synonym replacement (Feng et al., 2021). This technique replaces tokens within the text with synonymous words or phrases, with the aid of a knowledge base, such as WordNet[8]. Assuming the synonym does not change the semantics of the text, the synthetic document's labels should be the same. Synonym replacement with WordNet has been previously employed by Ollagnier and Williams (2020) in medical document classification. Their method randomly replaces a set number of non-stopwords per document with their synonyms. The relatively unrestricted choice of words, however, means the synonym replacement may not be applied to concepts of high interest – medical vocabulary. Schrempf et al. (2020) apply a focused form of document synthesis through the use of templates in radiology reports. These templates are used for augmenting concepts of interest, or replacing them with similar ones. UMLS-based synonym replacement has previously been used for DA in NER+L and sentence classification by Kang et al. (2021), employing random insertion, random swap, and random deletion, and UMLS-synonym replacement guided by the output of MetaMap.

We employ UMLS-synonym replacement DA similar to Kang et al. (2021) for the task of LMTC guided by more recent biomedical *NER+L* methods. We further propose a novel ontology-guided document synthesis turning relevant concepts into semantically adjacent concepts based on the ICD-9, with the expected label set being adjusted accordingly. The aim of this synthesis technique is to provide further training data specifically to few-shot and zero-shot labels.

2.5 Evaluation and Analysis for LMTC

LMTC tasks are evaluated using precision, recall and F1 score with micro and macro averaging, where macro-level metrics place equal weight on the performance on each label, disregarding the class imbalance. For the FS and ZS scenario precision and recall of the k highest predictions (@k), regardless of passing a fixed threshold tend to be employed. These measures compare exact match (intersection) between the prediction and gold standard sets ignoring the rich ontological structure and consider all errors equivalent. *Count-Preserving Hierarchical Evaluation* (CoPHE) (Falis et al., 2021) is a recently proposed evaluation metric involving the ontological structure to award partial credit to mispredictions occurring within the family of codes to which a gold-standard label belongs. Through the preservation of counts this method also considers over-/under-prediction within families of codes.

Beyond aggregate measures, to the best of our knowledge, label-specific analysis tools do not exist. Due to the weakly-labelled nature of LMTC tasks, confusion matrices are not a viable option. We introduce an analysis method akin to the confusion matrix suitable for LMTC.

3 Data

We employ the discharge summaries of MIMIC-III (Johnson et al., 2016) due to their common use in medical LMTC tasks. MIMIC-III is a multimodal medical dataset acquired from the intensive care units of the Beth Israel Deaconess Medical Center in Boston, Massachusetts between years 2001 and 2012. Access of the data was granted through PhysioNet[9] after completing the ethical training by the Collaborative Institutional Training Initiative program. The dataset is coded with ICD-9 codes on the document level. These labels do not perfectly represent the content of the text – MIMIC-III is significantly under-coded for specific conditions (Searle et al., 2020), and sometimes incorrect codes are assigned – *e.g.*, in the case of smoking status (Falis et al., 2019).

The data has been pre-processed and split following Mullenbach et al. (2018)'s procedures. The

[8] https://wordnet.princeton.edu/

[9] https://physionet.org/content/mimiciii/1.4/

label distribution within the dataset follows a big-head long-tail distribution. We divide the labels, similar to Rios and Kavuluru (2018), into three subsets according to their population size: Of the total 8,929 unique labels 4,351 appear in more than 5 documents within the training set; 4,341 at least once but at most 5 times (few-shot); and 237 labels do not appear in the training set, while existing in the development or test set (zero-shot). The training set consists of 47,719 documents.

4 Methods

Our methods include data augmentation and synthesis strategies based on the synonyms and adjacent concepts respectively, and an analytic tool for LMTC based on a set of assumptions to adapt confusion matrices with ontological structure.

4.1 Data Augmentation and Synthesis Strategies

We have attempted to enhance the training data with variety in the vocabulary and introduction of new codes in synthetic data. We applied two NER+L systems – SemEHR and MedCAT – to the training set. Unlike Searle et al. (2020) who sought to produce a silver standard by reconciling the output of NER+L methods with the gold standard, for the purpose of determining candidate codes for DA we chose to filter the NER+L outputs by intersecting them with the gold standard. While the gold standard may not capture all mentioned concepts, it may reflect local coding guidelines. As the NER+L systems label their outputs with CUIs, we translated these into ICD-9 using *PyMedTermino*.[10]

It should be noted that LMTC models, such as CAML, rely on pre-trained word2vec features with a static vocabulary – words unseen during pre-training will be considered *out-of-vocabulary* (OOV). This affects concepts that are unseen during training, such as rare diseases named after a person – *e.g.*, Munchausen's Syndrome (*301.51* in ICD-9). By introducing alternative names (augmentation) or new concepts (synthesis) we can also expand the relevant vocabulary, mitigating OOV.

4.1.1 Identity-Code Augmentation

We first created a synonym-replacement DA method in order to make the models more robust to variety. A medical concept can have several alternative names or surface forms including abbreviations – *e.g.*, an "acute myocardial infarction" can be referred to as "heart attack" or the abbreviation "MI". Through augmenting the text with synonyms we expose the model to alternative keywords representing existing concepts (already within the corpus or previously unseen), while leaving non-keyword context tokens untouched.

If an input document has any NER+L predictions matching the gold standard, their spans are identified. A synonym from PyMedTermino (derived from the UMLS, ICD-9, ICD-10, and SNOMED CT) is chosen at random, and replaced within the input text for each span. The augmented text is then added to the training set with the same gold standard labels as the original.

4.1.2 Adjacent-Code Synthesis

An additional form of *Document Synthesis* (DS), aimed at introducing new labels, can be produced by replacing mentions of a concept with an adjacent concept, rather than a synonym – *e.g.*, "stage 2 glaucoma" with "stage 3 glaucoma" – and updating the gold standard for the synthetic document accordingly. Where Identity-Code Augmentation aims to expose the model to alternative keywords to concepts pre-existing in the corpus without changing the code, the Adjacent-Code Synthesis replaces the code, exposing the model to the keyword of a different code – potentially one that is rare within the original training set (FS), or not appearing in it at all (ZS). This replacement leads to these keywords appearing in new contexts (those of the concepts they replace).

We chose to focus on "unspecified" codes assuming an "unspecified" label means all its mentions within are non-specific, while a single specified mention warrants a more specific version of the code in the new silver standard. This choice was made to address imperfections in the NER+L predictions – replacing a specified code would require replacement of all its mentions, some of which may not be identified by the NER+L method.

The outputs of SemEHR and MedCAT are processed as in the synonym-replacement DA. We considered a code to be unspecified if its description contained the string "unspecified" or "not otherwise specified", and with with "9" as the first or "0"/"1" as the second digit of the etiology. Of the 8,692 unique codes appearing in the training set 1,188 remained as viable "unspecified codes".

[10] https://pythonhosted.org/PyMedTermino/

This represents 14.74% of the total code population within the training set.

Replacement codes were identified depending on the etiology – double-digit unspecified codes can only be replaced by codes differing only in the final digit, while single-digit unspecified codes can be replaced with codes of the same category with any other etiology. Replacement codes were divided into three sets – frequent (>5), few-shot (at least one but up to 5), zero-shot (unseen) – based on their population in the training set. Only labels known to be within the MIMIC-III dataset were considered.

For a given document each viable unspecified code is first randomly converted into a specified candidate (with ZS and FS candidates being preferred). The mentions of the unspecified code are randomly replaced with mentions of the specified candidate. The resulting synthetic discharge summary is then added into the training set with the original gold standard code replaced with the candidate code. The pipeline for this DS procedure is presented in Figure 1.

4.1.3 Enriched Training Sets

We applied the synonym DA method in a single pass on the original training set for each NER+L method explored, resulting in the sets *SemEHR-DA* and *MedCAT-DA*. The adjacent label DS method was applied in two passes for each NER+L method. This was done to allow for multiple adjacent-code-synthetic alternatives per document. The resulting datasets are called *SemEHR-DS* and *MedCAT-DS*. *SemEHR-Both* and *MedCAT-Both* are the combinations of DA and DS datasets. All DA and DS datasets were combined with the *Baseline*, and deduplicated. The final sizes of the different training sets are presented in Table 1, including the number of unique codes within the frequent, few-shot, and zero-shot subsets. DA strategies increase populations of frequent and few-shot codes, leading to some few-shot codes becoming frequent (>5 occurrences in the training set). DS expands on this by also increasing populations of 13 zero-shot codes. The development set and test set were left unmodified.

4.2 Hierarchical Confusion Matrix

Confusion matrices are useful evaluation analysis tools in strongly-labelled scenarios (where individual predictions are associated with gold labels)(Tan et al., 2019, p. 138). A high misclassification between two classes indicates that, with respect to the model's parameters and the data, members of these classes are similar and difficult to distinguish.

Dataset	Size	Frequent	Few	Zero
Baseline	47,719	4,351	4,341	237
SemEHR-DA	66,559	4,818	3,874	237
MedCAT-DA	71,295	4,998	3,694	237
SemEHR-DS	74,851	5,167	3,538	224
MedCAT-DS	74,830	5,164	3,541	224
SemEHR-Both	93,690	5,446	3,259	224
MedCAT-Both	98,402	5,565	3,140	224

Table 1: Training set sizes (number of documents) and populations (number of unique codes) of the frequent, few-shot, and zero-shot subsets.

Confusion matrices can also support labels without a valid association, *e.g.*, a prediction on a span not present in the gold standard, by associating them with a special label indicating absence of the counterpart. This scenario represents over/under prediction.

The confusion matrix enables high-level error analysis beyond tracking precision and recall of the model. Such error analysis can be used in further model design, or serve as supplementary information for a deployed model.

In the weakly-labeled scenario, such as ICD-9 coding, both predictions and gold labels are presented on the document level as sets without associations between individual labels. If there is a mismatch between a predicted label and the gold standard, we cannot state with certainty that a predicted label, say, A.4 (*e.g.*, Alcohol abuse, continuous) was misclassified as gold standard label A.2 (*e.g.*, Alcohol abuse, episodic) or B.1 (*e.g.*, Chronic bronchitis), or whether the model overpredicted A.4, while underpredicting B.1 and A.2. We can, however, make assumptions based on the ontological structure associating mispredictions within code families – relating the A.4 prediction to the gold label A.2 rather than B.1.

The problem of analysing multi-label classification tasks and hierarchical label spaces with confusion matrices has attracted recent attention within the visualisation community (Görtler et al., 2021; Heydarian et al., 2022). Heydarian et al. (2022) propose an extension to the standard confusion matrix for multi-label classification in a non-hierarchical setting. Görtler et al. (2021) propose a method of analysis in a hierarchical multi-output setting, approaching high-dimensional confusion as a distribution.

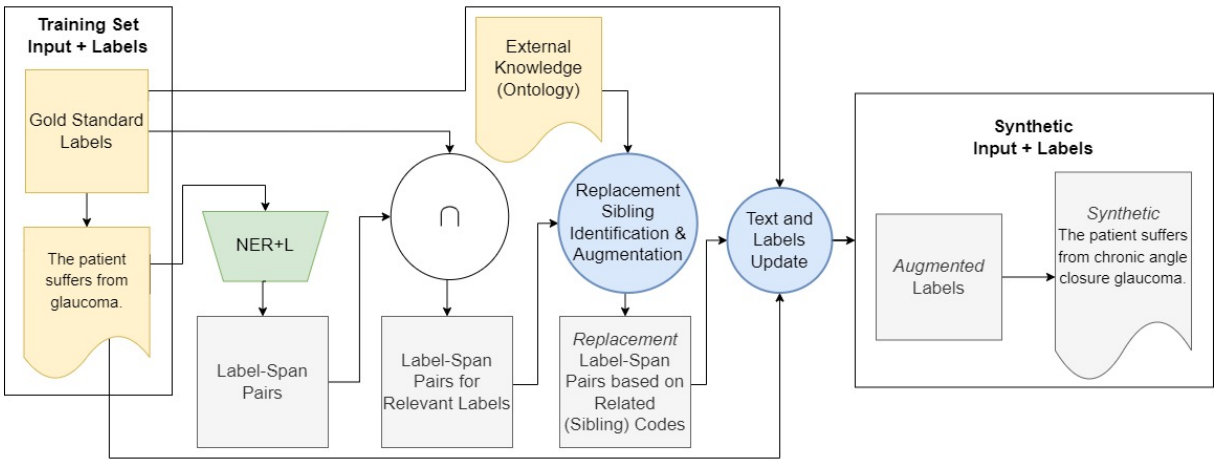

Figure 1: Ontology-Aided Document Synthesis pipeline. Yellow elements indicate data from human experts (input document, gold standard labels, ontology), gray elements indicate data which have machine learning somewhere in the creation process. The green element indicates pre-existing software, blue elements indicates software custom written for this method.

A co-occurrence matrix between predictions and gold labels indicates which predicted labels co-occur with particular gold standard labels, but is not fine-grained enough for error analysis. We propose the use of ontological structure to reduce the co-occurrence matrix into a simple weak hierarchical confusion matrix analysis method designed with the LMTC scenario in mind and apply it to ICD-9 coding. We further aggregate its results into performance metrics exploring proportion of errors based on their type.

4.2.1 Assumptions

Starting from a co-occurrence matrix between the predicted and gold standard sets of labels we apply three assumptions:

- *1-to-1 True Positive Correspondence*: If a label is present both in the prediction and gold-standard for a document, this is a True Positive (TP), and not considered for confusion.
- *Within-Family Confusion*: non-TP codes in the prediction are matched with non-TP codes in the gold standard within their respective code families (black cells in Figure 2 are ignored).
- *Out-Of-Family Scenario*: If in confusion matching a code from prediction/gold cannot be matched (no code from its family left to match), the code is associated with a special OOF code (see red the cell in Figure 2).

4.2.2 Use

While we are capable of visualising WHCMs (Appendix A.1) for each family, for the purposes of this publication we opt for aggregating results for all codes. In particular, we reduce the matrices into the following data given gold standard code: What proportion of the gold standard is correctly matched to its prediction, is confused within its family, and is in the OOF scenario. These three percentages sum up to 1. Furthermore, for each code we also track which code within its family (including OOF) is the most likely to be predicted, given the gold standard label. This information is used to determine if this most likely code matches the gold standard code. An example of this analysis can be found in Table A1 in the Appendix. We macro-average the correct-match/within-family confusion/OOF statistics and then provide the percentage of matches between preferred prediction given the gold standard. Note further analyses can be drawn conditioning on the predicted codes by applying the same procedures to the transpose of the original WHCM.

5 Experiments

We have applied MedCAT and SemEHR to the training set, producing candidate spans associated with CUIs. Post CUI-to-ICD-9 conversion, we have removed all candidates not matching the gold standard of their source discharge summary. We have produced augmented and synthetic data according to our description in sections 4.1.1 and 4.1.2 and combined them with the original train-

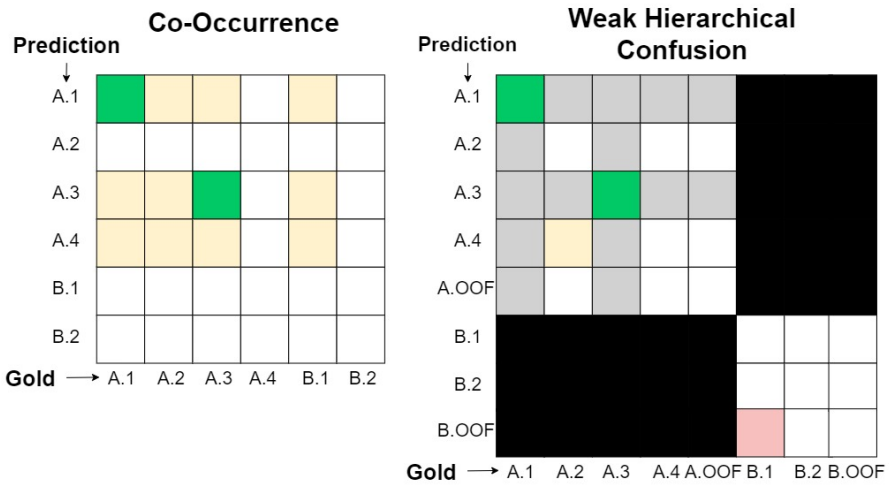

Figure 2: Left: A simple co-occurrence matrix between the prediction and gold standard labels for two label families for a single document. Labels A.1, A.3, and A.4 are predicted, while codes A.1, A.2, A.3, and B.1 are in the gold standard. Green cells indicate a match between the prediction and gold standard, yellow cells indicate a mismatch. Right: A weak hierarchical confusion matrix constructed from the co-occurrence matrix with the use of the three assumptions – Gray cells were eliminated via 1-to-1 correspondence, black cells were eliminated via within-family-confusion, red cells indicate the OOF scenario. The resulting confusion matrix indicates A.1 and A.3 being correctly predicted (green), B.1 being a false negative – an OOF (red), and the predicted code A.4 being confused with expected code A.2 (yellow).

ing set (dropping any duplicates) to produce enriched training sets as presented in section 4.1.3. We have further created a Baseline-like dataset of a similar size to our largest datasets – SemEHR-Both and MedCAT-Both – as a controlled experiment. This was done by concatenating two Baseline datasets (*2xBaseline*). Assuming a constant number of epochs, training on 2xBaseline corresponds to training on the Baseline for double the number of epochs.

We train CAML models based on the implementation of Chalkidis et al. (2019)[11] for 15 epochs on the training sets (Table 1). No few-shot/zero-shot model-side solution (such as the use of label embeddings as parameters) was applied. Each experiment used word embeddings of size 100 pre-trained on its respective training set according to Mullenbach et al. (2018)'s procedure. The development and test sets were the same across all experiments. The model weights with the best end-of-epoch validation F1 score were evaluated on the test set.

6 Results

For each experiment we report results averaged across 5 runs (Table 2), except the three largest (2xBaseline, SemEHR-Both, MedCAT-Both) for which a single run was conducted. We compare the performance on previously used metrics: Micro-F1 for all codes, and R@10 for few-shot and zero-shot codes. The codeset for few-shot and zero-shot codes is derived from the Baseline, and hence includes codes whose populations have increased in the DA, DS, and Both datasets. We further report hierarchical results (Mic-F1$_H$) according to CoPHE. Furthermore, we present macro-averaged aggregate measures conditioned on the *gold-standard labels* for all labels coming from WHCM – percentages of gold labels being being predicted correctly (Mac-Cor), being confused with a code within the same family (Mac-Conf), and being confused as OOF (underprediction – Mac-OOF). Finally, we track whether the prediction most often matched with the gold standard code, is the identity code itself (rather than a sibling or OOF) – if a correct prediction is more likely than any kind of misprediction. On a code level this is represented as a binary value (match or mismatch), and then can be macro-averaged to the metric Match. For our WHCM families we have used the ICD-9 tree as implemented in CoPHE aggregating on its parent level (code category). It should be noted, that our CAML baseline results underperform with respect

[11] https://github.com/iliaschalkidis/lmtc-emnlp2020

Dataset	Mic-F1	Mic-F1$_H$	R@10-Few	R@10-Zero	Mac-Cor	Mac-Conf	Mac-OOF	Match
Baseline	0.441	0.487	0.034	0.035	0.043	**0.055**	0.902	0.044
2xBaseline*	0.477	0.521	**0.093**	**0.075**	0.073	0.066	0.861	0.077
SemEHR-DA	0.469	0.514	0.055	0.034	0.062	0.062	0.876	0.063
MedCAT-DA	0.468	0.514	0.064	0.048	0.062	0.065	0.873	0.065
SemEHR-DS	0.471	0.518	0.051	0.055	0.067	0.065	0.869	0.069
MedCAT-DS	0.474	0.520	0.059	0.054	0.068	0.065	0.866	0.071
SemEHR-Both*	0.483	0.528	0.066	0.051	**0.079**	0.066	0.855	0.081
MedCAT-Both*	**0.486**	**0.532**	0.071	0.057	**0.079**	0.068	**0.853**	**0.083**

Table 2: Test-set performance of CAML models trained on the original training set (Baseline) versus training sets with synonym augmentation (SemEHR-DA, MeCAT-DA), and adjacent-code synthesis (SemEHR-DS, MedCAT-DS) averaged across 5 runs. Experiments on datasets marked with an asterisk (2xBaseline, SemEHR-Both and MedCAT-Both) have, due to time constraints, been conducted a single run each. Best performance for each metric is marked bold. Results are reported on the original test set. Zero and Few-shot codesets are based on the Baseline. The original development set is used for validation in all experiments.

to the official results of Mullenbach et al. (2018) (Micro-F1 of 0.53), due to our limited number of training epochs (while Mullenbach et al. (2018) ceases training after the precision@8 does not improve for 10 epochs).

All the proposed methods improve on the Baseline with regard to standard and hierarchical Micro-F1. Augmentation (DA) sets, while comparatively worse than Synthetic (DS) on R@10-Zero and standard and hierarchical Micro-F1, perform better on R@10-Few. This was to be expected as the DA methods provide little for the Zero codeset, while producing more of the labels in the Freq and Few codesets. Interestingly, SemEHR-DA performs on par with MedCAT-DA despite having a smaller training set. The combination of DA and DS methods (Both) report the best F1 results, with MedCAT-Both performing best in 5 our of the 8 reported metrics (including Mac-Cor, Mac-OOF and Match). Both of these methods' results are at least as good as those of 2xBaseline, which is of comparable size. The best R@10-Few and R@10-Zero performance was achieved by 2xBaseline, which corresponds to training the Baseline for twice as many epochs. While the different improvement of DA and DS in R@10-Few and R@10-zero implies our methods enhance these subsets, 2xBaseline dominating these metrics suggests that a better performance FS/ZS can be achieved with more training epochs. The difference between the standard and hierarchical (CoPHE) F1 scores remained largely the same, which implies partial errors were not addressed by these methods – this is further supported by the changes in Mac-OOF dominating compared to those of Mac-Conf. The lowest Mac-Conf was achieved by the original Baseline, but was coupled with a high OOF implying that this low confusion is mostly due to a higher proportion of codes not being predicted at all.

7 Conclusion and Discussion

The data enrichment methods have improved on the baseline showing potential in approaching the few-shot/zero-shot scenario through data, rather than the model. However, our approach relied on the use of external NER+L tools, whose predictions are imperfect, and may not be available for all domains of interest. Other avenues of finding relevant entities, *e.g.*, the attention outputs of LMTC models, should be explored in future work. While the data enrichment results are encouraging, further analysis on fully trained LMTC models is desirable. WHCM results point to a major issue with most false negatives coming from underprediction of a family, rather than within-family confusion. Further analysis should be conducted on false positives. The analysis from the WHCM tool can provide possible explanation of the errors of a model and may shed light on the design of more accurate models for LMTC.

Acknowledgements

This work is supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics, and the Health Data Research UK (HDR UK) National Text Analytics and Phenomics Projects. HD is supported by the Engineering and Physical Sciences Research Council (EPSRC, grant EP/V050869/1), Concur: Knowledge Base Construction and Curation.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on eu legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322.

Matúš Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. 2021. CoPHE: A count-preserving hierarchical evaluation metric in large-scale multi-label text classification. In *2021 Conference on Empirical Methods in Natural Language Processing*.

Matúš Falis, Maciej Pajak, Aneta Lisowska, Patrick Schrempf, Lucas Deckers, Shadia Mikhael, Sotirios Tsaftaris, and Alison O'Neil. 2019. Ontological attention ensembles for capturing semantic concepts in icd code prediction from clinical text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 168–177.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Genevieve Gorrell, Xingyi Song, and Angus Roberts. 2018. Bio-yodie: A named entity linking system for biomedical text. *arXiv preprint arXiv:1811.04860*.

Jochen Görtler, Fred Hohman, Dominik Moritz, Kanit Wongsuphasawat, Donghao Ren, Rahul Nair, Marc Kirchner, and Kayur Patel. 2021. Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. *arXiv preprint arXiv:2110.12536*.

Mohammadreza Heydarian, Thomas E Doyle, and Reza Samavi. 2022. MLCM: Multi-label confusion matrix. *IEEE Access*, 10:19083–19095.

Alistair EW 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(1):1–9.

Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. 2021. UMLS-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823.

Zeljko Kraljevic, Daniel Bean, Aurelie Mascio, Lukasz Roguski, Amos Folarin, Angus Roberts, Rebecca Bendayan, and Richard Dobson. 2019. Medcat–medical concept annotation tool. *arXiv preprint arXiv:1912.10166*.

Jueqing Lu, Lan Du, Ming Liu, and Joanna Dipnall. 2020. Multi-label few/zero-shot learning with knowledge aggregated from multiple label graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2935–2943.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *NAACL-HLT*.

Anaïs Ollagnier and Hywel TP Williams. 2020. Text augmentation techniques for clinical case classification. In *CLEF (Working Notes)*.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 3132. NIH Public Access.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Patrick Schrempf, Hannah Watson, Shadia Mikhael, Maciej Pajak, Matúš Falis, Aneta Lisowska, Keith W Muir, David Harris-Birtill, and Alison Q O'Neil. 2020. Paying per-label attention for multi-label extraction from radiology reports. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 277–289. Springer.

Thomas Searle, Zina Ibrahim, and Richard Dobson. 2020. Experimental evaluation and development of a silver-standard for the mimic-iii clinical coding dataset. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 76–85.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2019. *Introduction to Data Mining, (Second Edition, Global Edition)*. Pearson Education Limited, Harlow, United Kingdom.

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, et al. 2018. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association*, 25(5):530–537.

A Example Analysis Output

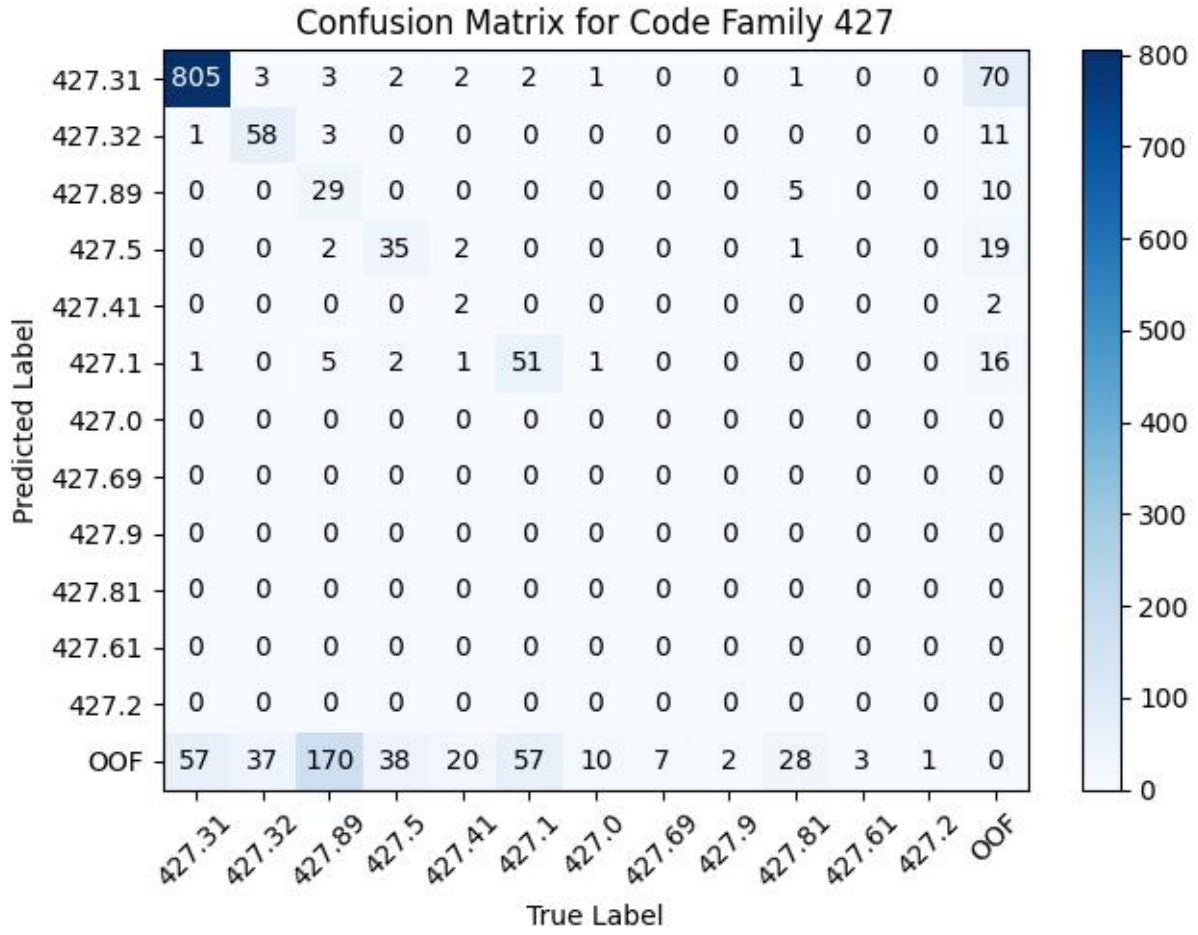

Figure A.1: An example of a WHCM for the code family 427 (Cardiac dysrhythmias). Codes that are predicted, are mostly predicted correctly (high-precision). Codes 427.31 (Atrial fibrillation) and 427.1 (Paroxysmal ventricular tachycardia) notably get confused with several of their siblings. While the model experienced some over-prediction, it suffered far more from under-prediction (low-recall).

Code	Identity #	Identity %	Preferred Prediction	Preferred Prediction #	Preferred Prediction %	OOF %	In-Family-Confusion %	Match
427.31	805	93.2	427.31	805	93.2	6.6	0.2	TRUE
427.32	58	59.2	427.32	58	59.2	37.8	3	TRUE
427.89	29	13.7	OOF	170	80.2	80.2	6.1	FALSE
427.5	35	45.5	OOF	38	49.4	49.4	5.1	FALSE
427.41	2	7.4	OOF	20	74.1	74.1	18.5	FALSE
427.1	51	46.4	OOF	57	51.8	51.8	1.8	FALSE
427.0	0	0	OOF	10	83.3	83.3	16.7	FALSE
427.69	0	0	OOF	7	100	100	0	FALSE
427.9	0	0	OOF	2	100	100	0	FALSE
427.81	0	0	OOF	28	80	80	20	FALSE
427.61	0	0	OOF	3	100	100	0	FALSE
427.2	0	0	OOF	1	100	100	0	FALSE

Table A1: An example of the output of the WHCM analysis tool for the *427* family of codes (Cardiac dysrhythmias, corresponding to the Figure A.1). There are 14 codes of this family present within MIMIC-III, with 12 appearing in the test set. Six of them have been correctly predicted at least once during the evaluation on the test set. Two of them (427.31, 427.32), are more likely to be predicted correctly than being confused within the family, or overpredicted (OOF). Four (427.89, 427.5, 427.41, 427.1) are predicted correctly at least once, but mostly suffer from underprediction (OOF). Six (427, 427.69, 427.9, 427.81, 427.61, 427.2) are never predicted correctly.