A Novel Framework Based on Medical Concept Driven Attention for Explainable Medical Code Prediction via External Knowledge

Anonymous ACL submission

Abstract

Medical code prediction from clinical notes aims at automatically associating medical codes with the clinical notes. Rare code problem, the medical codes with low occurrences, is prominent in medical code prediction. Recent studies employ deep neural networks and the external knowledge to tackle it. However, such approaches lack interpretability which is a vital issue in medical application. Moreover, due to the lengthy and noisy clinical notes, such approaches fail to achieve satisfactory results. Therefore, in this paper, we propose a novel framework based on medical concept driven attention to incorporate external knowledge for explainable medical code prediction. In specific, both the clinical notes and Wikipedia documents are aligned into topic space to extract medical concepts using topic modeling. Then, the medical concept-driven attention mechanism is applied to uncover the medical code related concepts which provide explanations for medical code prediction. Experimental results on the benchmark dataset show the superiority of the proposed framework over several state-of-the-art baselines.

1 Introduction

Medical codes, also known as ICD codes, are organized by International Classification of Diseases (ICD, recent versions are ICD-9 and ICD-10) taxonomies. Each medical code corresponds to a disease, procedure or sign, and so on. Medical codes can abstract away fine details of free-text clinical notes, which provide great convenience for analyzing clinical data directly (Shull, 2019; Bai and Vucetic, 2019). It is time consuming, costly and error-prone for manual medical coding due to the large menu of options (over 15,000 codes in ICD-9) and the complex lengthy clinical notes (Adams et al., 2002; Lang, 2007). Medical code prediction aims at automatically associating the relevant medical codes with the clinical notes.

ICD-9 code: 250
Disease name: Diabetes mellitus

Wikipedia Document:
Diabetes mellitus (DM), commonly referred to as **diabetes**, is a group of **metabolic disorders** in which there are **high blood sugar** levels over a prolonged period. Symptoms of high blood sugar include **frequent urination**, **increased thirst**, and **increased hunger**…
Diabetes is due to either the pancreas not producing enough **insulin**, or the cells of the body not responding properly to the insulin produced… Type 1 diabetes must be managed with **insulin injections**… Type 2 diabetes may be treated with medications such as **insulin sensitizers** with or without insulin.

Clinical Note:
Pt had been on **insulin** in past… Patient reports **frequent urination**, 10-15 times per day… You have been started on Lantus **Insulin** 25 units **injection** at bedtime. Please check your **blood sugars** twice daily, before breakfast and before nighttime insulin dose.

Figure 1: An example of a clinical note annotated with 3-digit ICD-9 code "250" and the corresponding Wikipedia document, where words in red are medical concept-indicative words which can be employed as evidences to infer medical codes.

Treating medical code prediction as a multi-label text classification problem, many machine learning based approaches have been proposed including Bayesian-based (Larkey and Croft, 1995) and Support Vector Machine based (Lita et al., 2008; Perotte et al., 2014). With the success of deep learning, many researchers propose neural networks with attention mechanism (Mullenbach et al., 2018; Li and Yu, 2020; Vu et al., 2020) to identify representative words in clinical notes and those with large weights serve as evidence for prediction. Li and Yu (2020) utilize convolutional neural networks with several fixed window sizes to capture various medical patterns, then identify representative ones through label attention mechanism.

Rare code problem, the medical codes with low occurrences, is prominent in medical code prediction. It was pointed out in (Bai and Vucetic, 2019) that among 942 3-digit ICD-9 codes occurring in the MIMIC-III database (the largest publicly-available medical database), the least common 437 codes account for only 1% of code oc-

currences. To tackle the rare code problem, Cao et al. (2020) leverage the code hierarchy and code co-occurrences information to aid predict rare codes. Vu et al. (2020) introduce a hierarchical joint learning architecture using the hierarchical relationships among codes to alleviate the rare code problem. Bai and Vucetic (2019) incorporate the external Wikipedia knowledge to enhance semantic information of the rare codes. The matching score between a clinical note and a medical code is calculated based on the code's related Wikipedia document.

However, most of the approaches mentioned above lack interpretability, which is vital for medical-related tasks. Moreover, most of these approaches fail to achieve satisfactory results because of the noisy and lengthy clinical notes (containing an average of 1,596 words). To address these challenges, we propose to explore latent medical concepts (including signs, symptoms, treatments, etc.) related to diseases, hidden in the clinical notes and Wikipedia knowledge. As shown in Figure 1, we can identify the informative medical concepts related to '*diabetes mellitus*', including signs '*high blood sugar*' and '*not enough insulin*', typical symptoms '*frequent urination, increased thirst, increased hunger*', and typical treatments '*insulin injection*' and '*insulin sensitizer*' based on the Wikipedia document describing ICD-9 code "*250*". Obviously, medical concepts mentioned above in clinical notes provide the effective evidences to infer disease '*diabetes mellitus*'. Moreover, based on the extracted medical concepts, the lengthy and noisy clinical notes can be alleviated.

Therefore, in this paper, we propose a novel framework based on medical concept driven attention (**MCDA**) to predict medical codes. Specifically, both the clinical notes and Wikipedia documents are fed as a whole corpus into the topic model to extract medical concepts. Both the Wikipedia documents and the clinical notes are represented as the distributions over the hidden topics (medical concepts) instead of the lengthy texts. Then, the medical concept-driven attention mechanism is applied, consisting of note-specific and label-specific concept-driven attention. On the one hand, the note-specific concept-driven attentions capture the salient medical concepts hidden in a specific clinical note. On the other hand, the label-specific concept-driven attentions focus on relevant medical concepts in a clinical note for each medical code. Experimental results show that the proposed framework outperforms a number of state-of-the-art models on a benchmark dataset.

The main contributions of this paper are listed as follows:

- A novel framework based on medical concept driven attention (**MCDA**) is proposed to predict medical codes. Moreover, the medical concept-driven attention mechanism, consisting of note-specific and label-specific concept-driven attention, is proposed to uncover the medical code related concepts hidden in the lengthy and noisy clinical notes. To the best of our knowledge, our work is the first attempt to explore latent medical concepts hidden in both the clinical notes and the external knowledge for explainable medical code prediction.

- Experimental results show that the proposed framework significantly outperforms several state-of-the-art models in all evaluation metrics. Moreover, it outperforms several state-of-the-art frameworks incorporating external knowledge in most evaluation metrics on the benchmark dataset.

2 Related work

Medical code prediction, also known as automatic ICD coding, is a challenging and important task in the limelight of medical informatics community.

Many traditional machine learning methods have been proposed including Bayesian-based (Larkey and Croft, 1995) and Support Vector Machine based (Lita et al., 2008; Perotte et al., 2014). Fueled by deep learning, many researchers have proven the effectiveness of convolutional neural network (CNN) and long short-term memory (LSTM) for medical code prediction. For example, Baumel et al. (2018) apply hierarchical attention networks for predicting medical codes. Mullenbach et al. (2018) propose a CNN with attention mechanism to capture relevant information in source text for each code. To find the specific evidence in the lengthy and noisy text for predicting accurately, researchers use CNN and variants (including multi-filter convolution, dilated convolution) with label attention mechanism to capture codes' relevant text patterns ($i.e.$ n-grams) in clinical notes (Mullenbach et al., 2018; Li and Yu, 2020; Ji et al., 2020). Vu et al. (2020) focus on label-specific words in notes via LSTM with customized label attention mechanism.

To tackle the rare code problem, different kinds of knowledge, such as label structure, label co-occurrence statistics, label descriptions and Wikipedia are employed. For example, the hierarchical tree structure of ICD-9 ontology is firstly exploited by (Perotte et al., 2014). Xie et al. (2019) employ graph convolutional network (GCN) to capture the hierarchical relationships among medical codes. Cao et al. (2020) construct a co-graph to incoporate code co-occurrence prior.

Instead of employing internal knowledge such label structure and label co-occurrence, some external knowledge are incorporated. Regarding label descriptions, Shi et al. (2017) apply character-aware neural network to match medical codes and clinical notes. Xie and Xing (2018) develop tree LSTM to use label descriptions. Zhou et al. (2021) train a teacher network with label descriptions and model the code co-occurrence through interactive shared attention. Regarding Wikipedia, Bai and Vucetic (2019) propose Knowledge Source Integration (KSI) framework to integrate Wikipedia documents describing medical codes during training of any baseline models. Compared with other external knowledge, Wikipedia knowledge is more informative and accessible.

Regarding incorporating the Wikipedia knowledge, the proposed approach is similar to KSI (Bai and Vucetic, 2019), but with the following significant differences: (1) we propose medical concept-driven attention to find note-specific and label-specific medical concepts in clinical notes as explainable evidences. While KSI simply calculates the matching score between a clinical note and a medical code's related Wikipedia document, unable to locate evidences in the context of clinical notes; (2) we make the most of Wikipedia knowledge through medical concepts, while KSI only considers the intersection of words in a clinical note and a Wikipedia document when predicting the corresponding medical code.

3 Methodology

3.1 Problem Setting

Given a collection of Q clinical notes denoted as $\mathcal{D} = \{d_1, d_2, \ldots, d_Q\}$. Each clinical note d_j consists of a sequence of words and is accompanied with a set of associated medical codes. We denote the size of medical code set $L = \{l_1, l_2, \ldots, l_{|L|}\}$ as $|L|$. In addition, we construct an external knowledge source $\mathcal{Z} = \{z_1, z_2, \ldots, z_{|L|}\}$ which consists of Wikipedia documents describing the medical codes. Each unique medical code l_i corresponds to a Wikipedia document z_i. Given a clinical note d_j, the goal is to predict the associated medical codes via the external knowledge source $\mathcal{Z}$, which can be treated as a multi-label text classification problem. Therefore, in the rest of the paper, medical codes are called labels for simplicity.

3.2 The Framework

The overall architecture of the proposed framework (**MCDA**) is shown in Figure 2, which consists of five parts:

(1) *Medical Concept Extraction Module* which extracts medical concepts from the clinical notes and Wikipedia documents; (2) *Embedding Layer* which includes word embeddings, medical concept embeddings and label embeddings; (3) *Encoder Layer* which includes backbone encoder and concept encoder; (4) *Concept-Driven Attention Layer* which calculates the note-specific and label-specific attention scores with the aid of medical concepts; (5) *Output Layer* which predicts the medical codes.

3.2.1 Medical Concept Extraction Module

The medical concepts are extracted via Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

At first, as the focuses and writing styles of Wikipedia documents and clinical notes are different, we pre-process both the Wikipedia documents and clinical notes. Words appearing in both the Wikipedia documents and clinical notes are retained.

Then, we feed the pre-processed $\mathcal{D}$ and $\mathcal{Z}$ as a whole corpus with vocabulary size V^c, into LDA to generate medical concepts. The granularity of the extracted medical concepts is controlled by the predefined K, the number of medical concepts.

Based on LDA, we obtain overall medical concept-word distribution matrix $\boldsymbol{C} \in \mathbb{R}^{K \times |V^c|}$. For the single clinical note d_j, the note-concept distribution $\boldsymbol{p}_j = (p_{j1}, p_{j2}, \ldots, p_{jK})$ represents the probability of the clinical note over each medical concept. Likewise, for a single Wikipedia document z_i (corresponding to label l_i), the label-concept distribution $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iK})$ represents the probability of the label over each medical concept. Thereby, the labels-concept distribution matrix is represented as $\boldsymbol{W} \in \mathbb{R}^{|L| \times K}$.

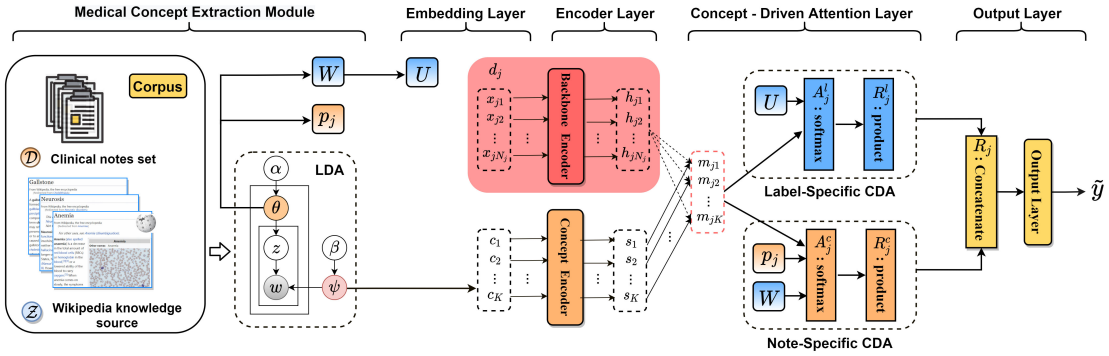

Figure 2: The architecture of the proposed **MCDA** Framework. $\mathcal{Z}$ and $\mathcal{D}$ denote the Wikipedia knowledge source and clinical notes set respectively. Label-Specific CDA represents label-specific concept-driven attention, and Note-Specific CDA represents note-specific concept-driven attention. It is worth noting that Backbone Encoder can be any neural encoders.

3.2.2 Embedding Layer

Embedding layer contains word embeddings, medical concept embeddings and label embeddings.

As for word embeddings, a clinical note d_j with N_j words is represented as $d_j = \{x_{j1}, x_{j2}, \ldots, x_{jN_j}\}$ using pre-trained word embeddings.

As for medical concept embeddings, the k^{th} medical concept's embedding $\boldsymbol{c}_k$ can be obtained from the overall medical concept-word distribution matrix $\boldsymbol{C}$.

With respect to the label embeddings matrix $\boldsymbol{U} \in \mathbb{R}^{|L| \times K}$, we use the labels-concept distribution matrix $\boldsymbol{W}$ as the initialization of $\boldsymbol{U}$ since LDA can capture the medical concepts information hidden in labels and implicitly model correlations between labels and clinical notes by projecting them into the same feature space.

3.2.3 Encoder Layer

Encoder layer contains both the backbone encoder and the concept encoder.

As for the backbone encoder, theoretically it can be any neural encoders, such as CNN based encoders, RNN based encoders or Transformer (Vaswani et al., 2017) based encoders. Given the clinical note $d_j = \{x_{j1}, x_{j2}, \cdots, x_{jN_j}\}$, the hidden state of each word is generated by the backbone encoder. Thereby, the clinical note d_j can be encoded as $\boldsymbol{h}_j = (h_{j1}, h_{j2}, \ldots, h_{jN_j})^\top \in \mathbb{R}^{N_j \times t}$, where t is the dimension of the hidden state.

As for the concept encoder, concept representations are produced by a fully connected layer followed by ReLU activation function taking the medical concept-word distribution matrix $\boldsymbol{C}$ as inputs. Hence, each concept representation $\boldsymbol{s}_k \in \mathbb{R}^t$ is obtained according to the medical concept embedding $\boldsymbol{c}_k$, $k \in \{1, 2, 3, \ldots, K\}$.

3.2.4 Concept-Driven Attention Layer

Not all words in the clinical note contribute equally to the decision of medical diagnosis. Moreover, not all medical concepts hidden in the clinical note contribute equally for medical code prediction. Therefore, attention weights are utilized to enhance clinical note representations according to both word representations and concept representations. We aggregate the representations of medical concepts-indicative words to form the clinical note representation.

Given the k^{th} concept representation $\boldsymbol{s}_k$, we can measure the interaction of words in the clinical note d_j and the medical concept by an attention weight vector $\boldsymbol{m}_{jk}$, which can be computed as the inner product of $\boldsymbol{s}_k$ and $\boldsymbol{\varphi}_j$ as follows,

$$\begin{aligned} \boldsymbol{\varphi}_j &= \tanh\left(\boldsymbol{h}_j \boldsymbol{W}^c + \boldsymbol{b}^c\right) \\ \boldsymbol{m}_{jk} &= \boldsymbol{\varphi}_j \, \boldsymbol{s}_k \end{aligned} \tag{1}$$

where $\boldsymbol{h}_j = \left(h_{j1}, h_{j2}, \ldots, h_{jN_j}\right)^\top$ stands for the combination of all hidden states of words in the clinical note d_j, $\boldsymbol{W}^c \in \mathbb{R}^{t \times t}$ and $\boldsymbol{b}^c \in \mathbb{R}^t$ are trainable parameters, $\boldsymbol{\varphi}_j = (\varphi_{j1}, \varphi_{j2}, \ldots, \varphi_{jN_j})$ refers to $\boldsymbol{h}_j$. The attention weight vector $\boldsymbol{m}_{jk}$ indicates how much attention the k^{th} medical concept pays to each word of the clinical note d_j.

Then, we propose two kinds of attention mechanisms including *Note-Specific Concept-Driven At-*

tention and *Label-Specific Concept-Driven Attention* based on $\boldsymbol{m}_{jk}$ in (1).

Note-Specific Concept-Driven Attention:

The note-specific concept-driven attention mechanism is employed to attend to note-specific medical concept words distributed in the clinical note. It leverages the note-specific medical concept information based on the note-concept distribution $\boldsymbol{p}_j = (p_{j1}, p_{j2}, \ldots, p_{jK})$ with each dimension representing the level of prominence of the corresponding medical concept occurred in the clinical note d_j. Then, it leverages the label-concept distribution matrix $\boldsymbol{W} \in \mathbb{R}^{|L| \times K}$ to generate an attention weight vector for each label. Given the clinical note d_j, for the i^{th} label, the note-specific concept-driven attention is calculated as follows,

$$\begin{aligned} \boldsymbol{a}^c_{ji} &= \text{softmax}\Big(\sum_{k=1}^{K} \boldsymbol{m}_{jk} \boldsymbol{p}_{jk} \boldsymbol{W}_{ik}\Big) \\ \boldsymbol{r}^c_{ji} &= \left(\boldsymbol{a}^c_{ji}\right)^\top \boldsymbol{h}_j \end{aligned} \tag{2}$$

for the i^{th} label, $\boldsymbol{a}^c_{ji}$ stands for the attention weight after incorporating the note-concept distribution $\boldsymbol{p}_j$ along with the label-concept distribution $\boldsymbol{W}_i$, to discover medical concept keywords that a single clinical note concerns for the specific label. The final note-specific concept-driven clinical note representation matrix $\boldsymbol{R}^c_j = \left(r^c_{j1}; r^c_{j2}; \ldots; r^c_{j|L|}\right)$ is constructed with the sum of hidden states $\boldsymbol{h}_j$ weighted by $\boldsymbol{A}^c_j = \left(\boldsymbol{a}^c_{j1}, \boldsymbol{a}^c_{j2}, \ldots, \boldsymbol{a}^c_{j|L|}\right)$. Each i^{th} row $\boldsymbol{r}^c_{ji}$ of the matrix $\boldsymbol{R}^c_j$ is the note-specific clinical note representation regarding the i^{th} label.

Label-Specific Concept-Driven Attention:

The label-specific concept-driven attention mechanism is proposed to capture label relevant medical concept words hidden in clinical notes using label embeddings. Given the clinical note d_j, for the i^{th} label, label-specific concept-driven attention is calculated as follows,

$$\begin{aligned} \boldsymbol{a}^l_{ji} &= \text{softmax}\Big(\sum_{k=1}^{K} \boldsymbol{m}_{jk} \boldsymbol{U}_{ik}\Big) \\ \boldsymbol{r}^l_{ji} &= \left(\boldsymbol{a}^l_{ji}\right)^\top \boldsymbol{h}_j \end{aligned} \tag{3}$$

We construct the label-specific clinical note representation matrix $\boldsymbol{R}^l_j = \left(\boldsymbol{r}^l_{j1}; \boldsymbol{r}^l_{j2}; \ldots; \boldsymbol{r}^l_{j|L|}\right)$ with the sum of hidden states $\boldsymbol{h}_j$ weighted by $\boldsymbol{A}^l_j = \left(\boldsymbol{a}^l_{j1}, \boldsymbol{a}^l_{j2}, \ldots, \boldsymbol{a}^l_{j|L|}\right)$.

Frequency range	Number of medical codes	Percentage of code occurrences
1-10	80	0.1%
11-50	73	0.6%
51-100	25	0.6%
101-500	82	6.7%
>500	84	92.0%

Table 1: Label frequency distribution

3.2.5 Output Layer

At last, we concatenate both representations calculated by note-specific and label-specific concept-driven attention to obtain final representation matrix $\boldsymbol{R}_j = [\boldsymbol{R}^c_j, \boldsymbol{R}^l_j]$ of clinical note d_j. $\boldsymbol{R}_j$ is then fed to a multi-layer perceptron (MLP) followed by the Sigmoid activation function for predicting all associated medical codes. This process can be formalized as follow,

$$\tilde{\boldsymbol{y}} = Sigmoid(MLP(\boldsymbol{R}_j)) \tag{4}$$

The training objective is to minimize the binary cross entropy loss between the prediction score $\tilde{\boldsymbol{y}}$ and the target $\boldsymbol{y}$:

$$Loss = -\sum_{i=1}^{|L|} \{y_i \log(\tilde{y}_i) + (1 - y_i) \log(1 - \tilde{y}_i)\} \tag{5}$$

4 Experiments

In this section, we describe the datasets, evaluation metrics, baselines and implementation details, before discussing the experimental results.

4.1 Dataset

The dataset is constructed based on clinical notes in MIMIC-III dataset and Wikipedia documents of ICD-9 diagnosis codes following the same way in (Bai and Vucetic, 2019). There are 52,722 condensed clinical notes in MIMIC-III (Johnson et al., 2016) dataset. On average, each note has 1,596 words. All medical codes are grouped by their first three digits. A subset of 344 medical codes is kept where each medical code has the corresponding Wikipedia document. On average, each Wikipedia document has 1,058 words. The whole word vocabulary contains 60,968 unique words, out of which only 12,173 can be found in clinical notes. It can be deduced that both the clinic notes and Wikipedia documents share significantly different word distributions.

Model	AUC		F1		Top-10 recall
	Macro	Micro	Macro	Micro	
CAML (Mullenbach et al., 2018)	0.855	0.978	0.257	0.656	0.806
+KSI (Bai and Vucetic, 2019)	0.891	0.980	0.285	0.659	0.814
+MCDA (ours)	**0.894 ± 0.004**	**0.982 ± 0.001**	**0.300 ± 0.010**	**0.679 ± 0.001**	**0.828 ± 0.001**
MultiResCNN (Li and Yu, 2020)	0.864 ± 0.008	0.980 ± 0.001	0.301 ± 0.011	0.673 ± 0.002	0.823 ± 0.001
+KSI (Bai and Vucetic, 2019)	**0.892 ± 0.005**	**0.982 ± 0.001**	**0.320 ± 0.010**	0.682 ± 0.002	**0.830 ± 0.001**
+MCDA (ours)	0.883 ± 0.005	**0.982 ± 0.001**	0.284 ± 0.008	**0.684 ± 0.004**	0.827 ± 0.002
DCAN (Ji et al., 2020)	0.847 ± 0.008	0.980 ± 0.001	0.260 ± 0.008	0.665 ± 0.002	0.822 ± 0.001
+KSI (Bai and Vucetic, 2019)	0.880 ± 0.005	0.981 ± 0.002	0.302 ± 0.011	0.679 ± 0.003	**0.831** ± 0.002
+MCDA (ours)	**0.898 ± 0.006**	**0.982 ± 0.001**	**0.311 ± 0.008**	**0.684 ± 0.001**	**0.831 ± 0.001**
LAAT (Vu et al., 2020)	0.899 ± 0.006	0.983 ± 0.001	0.342 ± 0.010	0.687 ± 0.003	0.835 ± 0.002
+KSI (Bai and Vucetic, 2019)	0.908 ± 0.003	**0.984 ± 0.001**	0.352 ± 0.010	0.690 ± 0.003	0.837 ± 0.001
+MCDA (ours)	**0.918 ± 0.006**	**0.984 ± 0.001**	**0.362 ± 0.008**	**0.702 ± 0.003**	**0.844 ± 0.002**

Table 2: Performance comparisons among several baselines and their counterparts under KSI framework and the proposed MCDA framework. We run all approaches 10 times with the same hyper-parameters using different random seeds except CAML and CAML+KSI, statistics of which are from the source paper. We report the $mean \pm standard\ deviation$ for each approach.

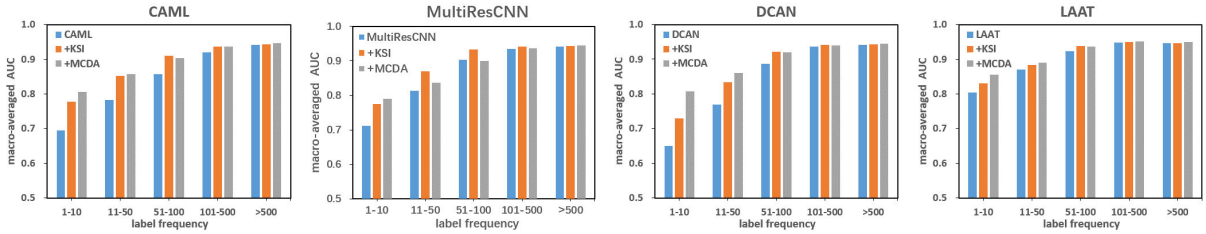

Figure 3: Macro-averaged AUC by label frequency group for CAML, MultiResCNN, DCAN and LAAT. x-axis denotes the label frequency group and y-axis denotes the macro-averaged AUC for each group.

Model	AUC		F1		Top-10 recall
	Macro	Micro	Macro	Micro	
LAAT+MCDA	**0.918**	**0.984**	**0.362**	**0.702**	**0.844**
w/o medical concept	0.899	0.983	0.342	0.687	0.835
w/o label-specific	0.872	0.974	0.223	0.630	0.772
w/o note-specific	0.904	0.983	0.342	0.686	0.833
w/o note-concept	0.915	0.983	0.350	0.698	0.842
w/o label-concept	0.912	**0.984**	0.345	0.698	0.843

Table 3: Ablation results.

4.2 Evaluation Metrics

We evaluate the proposed method using micro and macro AUC, F1 metrics and Top-10 recall following the same way in (Bai and Vucetic, 2019). As shown in Table 1, medical codes in the dataset is highly imbalanced, the most common 84 codes account for 92% of all code occurrences. We employ macro metrics to emphasize on rare code prediction.

4.3 Baselines

We choose four state-of-the-art models as the baselines, which employ the label attention mechanism over neural word encoders. Moreover, for fair comparison, all the baselines (as backbone encoder) are combined with KSI framework (Bai and Vucetic, 2019) and the proposed framework (MCDA) respectively to incorporate Wikipedia knowledge. Details of the baselines are described as follows:

KSI: Bai and Vucetic (2019) proposed the Knowledge Source Integration framework to integrate the Wikipedia knowledge. It can be combined with some medical code prediction baselines.

CAML: Mullenbach et al. (2018) proposed the convolutional attention network, which learns attention distribution for each medical code.

MultiResCNN: Li and Yu (2020) utilized the multi-filter convolutional layer to capture variable medical patterns and residual block to enlarge model's receptive field, incorporating the label attention mechanism to generate label-aware representations.

DCAN: Ji et al. (2020) integrated dilated convolutions and residual connections to capture complex medical patterns and also incorporated label attention mechanism.

LAAT: Vu et al. (2020) proposed the customized label attention model to learn attention distributions over BiLSTM encoding hidden states for each medical code.

4.4 Implementation Details

We use word2vec (Mikolov et al., 2013) to pre-train word embeddings with the size of 100 from clinical notes. The number of extracted medical concepts K is set to 100. We utilized default Adam optimizer (Kingma and Ba, 2014) to minimize the loss function. Regarding the training of the baseline models, we perform a grid search over hyperparameters according to their default parameter setting.

4.5 Results

Table 2 shows the performance comparisons among baselines and their counterparts under KSI framework and the proposed MCDA framework.

Overall, it can be observed that by employing the KSI or MCDA framework, the performances of all the baseline models are improved, which shows the effectiveness and necessity of incorporating the Wikipedia knowledge for medical codes. It is worth noting that, compared with KSI, MCDA improves baselines more significantly in most metrics. The great improvement of top-10 recall demonstrates the effectiveness of the proposed framework in recommending relevant medical codes. It is noteworthy that MCDA outperforms most baselines with a larger margin than KSI (except MultiResCNN) on the macro metric. As the performance on the macro metric shows how well the rare codes problem is handled, we can deduce that MCDA captures precisely the representations of labels and notes based on the medical concept from the external Wikipedia documents, which is crucial for rare codes.

To further validate this deduction, we divide medical codes into 5 groups based on their frequencies in the dataset as shown in Table 1: [1, 10], [11, 50], [51, 100], [101, 500] and [500, $+\infty$). We calculate macro-averaged AUC of each medical code group for all baselines and their counterparts under KSI framework and our MCDA framework. The results are summarized in Figure 3. It can be observed that both KSI and MCDA bring major improvements of AUC on the least common [1-10] and [11-50] group. For DCAN and CAML, MCDA improves much more than KSI on [1-10] group, 7.8% of DCAN and 2.8% of CAML. For the best baseline LAAT, MCDA improves 5.1% on [1-10] group and 2.1% on [11-50] group, which is better than 2.6% and 1.4% of KSI. The results demonstrate the benefit of incorporating medical concept driven attention than KSI in handling rare codes.

For MultiResCNN, though MCDA brings improvements on [1-10] group, it performs worse on [11-50], [51-100] group and the overall dataset. The possible reason is that, MultiResCNN concatenates outputs from 6 kernels with different sizes to generate hidden state h_i. Therefore, h_i is the simple concatenation of 6 n-grams' hidden states, not the hidden state of the ith word (or n-gram) in other baselines. Actually when the number of kernels decrease to 1, MultiResCNN degrades to CAML. It performs better on macro metrics which indicates that MultiResCNN is unsuitable for MCDA framework.

In addition, we also try Transformer (Vaswani et al., 2017) and pre-trained BERT (Devlin et al., 2018) as backbone encoder. However, no Transformer based models work well in this task mainly due to excessively long text. This conclusion is also reported in (Li and Yu, 2020), (Ji et al., 2020) and (Pascual et al., 2021).

4.6 Ablation Study

To further evaluate the effectiveness of each component, we conduct some ablation experiments on LAAT+MCDAM. The ablation results are shown in Table 3. It can be observed that:

Effectiveness of Medical Concept Without medical concept (w/o medical concept in Table 3), medical concept-driven attention degrades to the label attention mechanism proposed in LAAT. The performance drops on all metrics, especially on macro metrics, indicating a significant reduction in the ability to predict rare codes.

Effectiveness of Label-Specific Concept-Driven Attention When discarding the label-specific concept-driven attention (w/o label-specific in Table 3), the performance drops dramatically on all metrics, especially on F1 metric. It shows the effectiveness of label-specific concept driven attention in capturing desired labels' relevant information in lengthy and noisy clinical notes.

Effectiveness of Note-Specific Concept-Driven Attention When discarding the note-specific concept-driven attention (w/o note-specific in Table 3), the performance drops obviously. To further investigate the contribution of note-concept distribution p_j and labels-concept distribution matrix W, we remove them separately. Both the performances drop slightly. It can be concluded that they both are complementary for note-specific concept-driven attention.

Concept. (a) **Concept. (b)**

Figure 4: Word clouds of some medical concepts.

Clinical Note: The patient was taken for an ERCP the day after admission, where he was found to have a **stone** in the **common** bile duct treated with sphincterotomy and **stone** extraction … he then appeared to bleed from his **sphincterotomy**. His initial hematocrit was 40 and it then fell to 25. He received 2 units of transfusion let him go home with a planned return for an elective **cholecystectomy**. **-- (LAAT)**
Intersection: number liver sex **the** broad within **common** last **hematocrit** first several name **bleed** treat following total … **Intersection:** number liver upper **stone** sphincterotomy right **pain** within **bile duct** ercp surgery history surrounding acute … **-- (LAAT+KSI)**
Clinical Note: The patient was taken for an **ERCP** the day after admission, where he was found to have a stone in the common bile duct treated with **sphincterotomy** and stone extraction … he then appeared to bleed from his sphincterotomy. His initial hematocrit was 40 and it then **fell** to 25. He received 2 units of **transfusion**. **-- (LAAT+MCDA)**
Medical Codes: 285 (anemia); 574 (gallstone);

Figure 5: The attention distribution visualization over a clinical note with two medical codes for LAAT and its counterparts under KSI and MCDA framework. Regarding LAAT, the words in bold represent highly weighted ones by its label attention. Regarding KSI, the bold words are extracted keywords in the intersection with high attention weights. Regarding MCDA, the words in bold represent highly weighted ones by note-specific attention, while the words with underlines are highly weighted ones by label-specific attention.

5 Discussion

Medical Concept Visualization

We randomly select two medical concepts with their top-20 weighted words. The corresponding word clouds are shown in Figure 4, where the size of a word is proportional to its assigned weight. **Concept (a)** is a medical concept about disease 'diarrhea' accompanied with symptoms including 'vomiting', 'nausea', 'chills', 'pain', etc. **Concept (b)** is diseases of 'biliary and pancreatic' which also includes 'pancreatitis', 'ercp' (a medical test technique), 'bile duct' (organ), etc. These medical concepts can aggregate medical information including diseases, symptoms, diseased organs, treatments and so on, which can be used to describe clinical notes concisely and provide interpretability.

Case Study of Interpretability

To further explore interpretability of the proposed approach, the attention distribution visualization over a clinical note for LAAT and its counterparts under KSI and our MCDA is shown in Figure 5.

It can be observed that LAAT (Vu et al., 2020) with customized label attention mechanism only captures scattered label-related words like **'stone'** and **'cholecystectomy'** for inferring **'gallstone'**, while it fails to find valid relevant evidence for inferring **'anemia'**. KSI (Bai and Vucetic, 2019) can additionally aid LAAT to find out keywords relevant to the medical codes in the intersection of the corresponding Wikipedia document and the clinical note. However, KSI represents the intersection as a binary vector encoding the presence of words, which inevitably causes a great loss of information in the clinical note, and is unable to aid LAAT locate evidence in the context of the clinical notes for predicting corresponding medical code.

In contrast, MCDA's label-specific concept-driven attention guides LAAT discover the sign **'stone'** in **'bile duct'** which directly leads to **'gallstone'**. Moreover, based on note-specific concept-driven attention, some important medical concepts are retrieved and focused, such as **'ERCP'** (a medical test technique) and **'sphincterotomy'** (a specific surgery) which are strongly related to **'gallstone'**. Regarding medical code **'anemia'**, based on label-specific concept-driven attention, medical concepts **'bleed'** and **'hematocrit'** related to **'anemia'** are captured, and the medical sign **'hematocrit fell'** and treatment **'transfusion'** which can infer disease **'anemia'** are found based on note-specific concept-driven attention. Therefore, through medical concept-driven attention mechanism, different kinds of medical concepts are focused which provide more interpretability.

6 Conclusions

We have presented a novel framework based on medical concept driven attention for explainable medical code prediction from clinical notes. To the best of our knowledge, our work is the first attempt to uncover and explore latent medical concepts guided by the external knowledge while medical concept-indicative words serve as the evidences for explainable medical code prediction. Experimental results show that MCDA improves significantly several state-of-the-art models in most evaluation metrics on the benchmark dataset. In future, more Wikipedia documents will be incorporated and other ways of incorporating will be explored to promote medical code prediction task.

References

Diane L Adams, Helen Norman, and Valentine J Burroughs. 2002. Addressing medical coding and billing part ii: a strategy for achieving compliance. a risk management approach for reducing coding and billing errors. *Journal of the National Medical Association*, 94(6):430.

Tian Bai and Slobodan Vucetic. 2019. Improving medical code prediction from clinical text via incorporating online knowledge sources. In *The World Wide Web Conference*, pages 72–82.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Hypercore: Hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2020. Dilated convolutional attention network for medical code assignment from clinical text. *arXiv preprint arXiv:2009.14578*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, 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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dee Lang. 2007. Consultant report-natural language processing in the health care industry. *Cincinnati Children's Hospital Medical Center, Winter*, 6.

Leah S Larkey and W Bruce Croft. 1995. Automatic assignment of icd9 codes to discharge summaries. Technical report, Technical report, University of Massachusetts at Amherst, Amherst, MA.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8180–8187.

Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. 2008. Large scale diagnostic code classification for medical patient records. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

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

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards BERT-based automatic ICD coding: Limitations and opportunities. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 54–63, Online. Association for Computational Linguistics.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*.

Jessica Germaine Shull. 2019. Digital health and the state of interoperable electronic health records. *JMIR medical informatics*, 7(4):e12712.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. *arXiv preprint arXiv:2007.06351*.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated icd coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076.

Xiancheng Xie, Yun Xiong, Philip S Yu, and Yangyong Zhu. 2019. Ehr coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 649–658.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic icd coding via interactive shared

representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957.