Clinical-Coder: Assigning Interpretable ICD-10 Codes to Chinese Clinical Notes

Pengfei Cao[*1,2], **Chenwei Yan**[*3],
Xiangling Fu[3], **Yubo Chen**[1,2], **Kang Liu**[1,2], **Jun Zhao**[1,2], **Shengping Liu**[4], **Weifeng Chong**[4],
[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China
[3] Key Laboratory of Trustworthy Distributed Computing and Service(BUPT),
Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China
[4] Beijing Unisound Information Technology Co., Ltd, Beijing, 100028, China
{pengfei.cao, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn, {2013213347, fuxiangling}@bupt.edu.cn,
{liushengping, chongweifeng}@unisound.com

Abstract

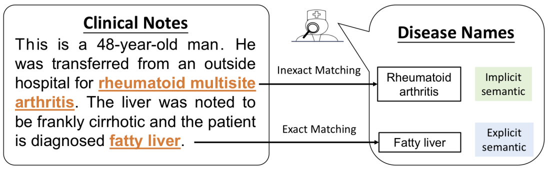
In this paper, we introduce Clinical-Coder, an online system aiming to assign ICD codes to Chinese clinical notes. ICD coding has been a research hotspot of clinical medicine, but the interpretability of prediction hinders its practical application. We exploit a **D**ilated **C**onvolutional **A**ttention network with **N**-gram **M**atching Mechanism (DCANM) to capture semantic features for non-continuous words and continuous n-gram words, concentrating on explaining the reason why each ICD code to be predicted. The experiments demonstrate that our approach is effective and that our system is able to provide supporting information in clinical decision making.

1 Introduction

International Classification of Disease (ICD) is the diagnostic classification standard in the field of clinical medicine, which assigns unique code to each disease. The popularization of ICD codes immensely promotes the information sharing and clinical research of disease worldwide and has a positive influence on health condition research, insurance claims, morbidity and mortality statistics (Shi et al., 2017). Therefore, ICD coding – which assigns proper ICD codes to a clinical note – has drawn much attention.

It is always that ICD coding relies on the manual work of professional staff. The manual coding is very error-prone and time-consuming since the continuous updating version of ICD codes results in a substantial increase in code numbers. The number of ICD-10 codes is up to 72,184, more than five times the previous version (i.e., ICD-9). It allows for more detailed classifications of patients' conditions, injuries, and diseases. However, there is no doubt that the increased granularity increases the difficulty of manual coding.

Existing studies came up with several approaches of automatic coding prediction to replace the repetitive manual work, from the traditional machine learning methods (Perotte et al., 2013; Koopman et al., 2015), to neural network methods (Shi et al., 2017; Yu et al., 2019). Although these methods achieve great success, they are still confronted with a critical challenge, which is the interpretability of predicted codes. Explainable model and results are essential for clinical medicine decision making (Mullenbach et al., 2018). Thus, the practical approach is supposed to predict correct codes and simultaneously give the reason why each code is predicted.



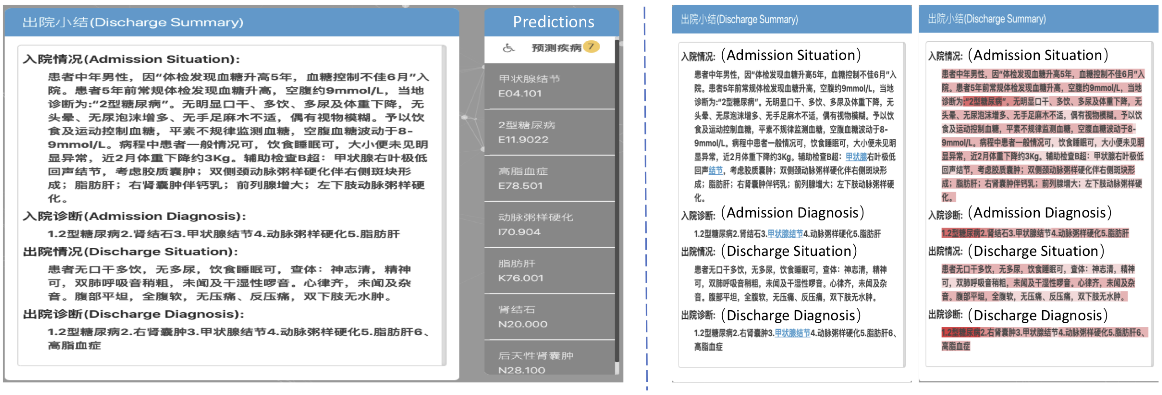
Clinical Notes

This is a 48-year-old man. He was transferred from an outside hospital for rheumatoid multisite arthritis. The liver was noted to be frankly cirrhotic and the patient is diagnosed fatty liver.

Figure 1: Two kind of semantic phenomenon: explicit semantic features and implicit semantic features.

In this paper, we try to provide the interpretability of predictions from a semantic perspective. It is a phenomenon that the exact disease names or similar expressions of disease names often appear in the discharge summary. For example, as shown in Figure 1, the exact matching with disease name such as "fatty liver" is a direct evidence of inference. We call the continuous consistent words as explicit semantic features. Moreover, the inexact matching such as "rheumatoid multisite arthritis" is also very useful to predict the codes and should be taken into consideration. We refer to the non-continuous

[*] co-first authors, they contributed equally to this work



出院小结(Discharge Summary)

入院情况(Admission Situation):

患者中年男性，因“体检发现血糖升高5年，血糖控制不佳6月”入院。患者5年前常规体检发现血糖升高，空腹约9mmol/L，当地诊断为:“2型糖尿病”。无明显口干、多饮、多尿及体重下降，无头晕、无尿泡沫增多、无手足麻木不适，偶有视物模糊。予以饮食及运动控制血糖，平素不规律监测血糖，空腹血糖波动于8-9mmol/L。病程中患者一般情况可，饮食睡眠可，大小便未见明显异常，近2月体重下降约3Kg。辅助检查B超：甲状腺右叶极低回声结节，考虑胶质囊肿；双侧颈动脉粥样硬化伴右侧斑块形成；脂肪肝；右肾囊肿伴钙乳；前列腺增大；左下肢动脉粥样硬化。

入院诊断(Admission Diagnosis):

1.2型糖尿病2.肾结石3.甲状腺结节4.动脉粥样硬化5.脂肪肝

出院情况(Discharge Situation):

患者无口干多饮，无多尿，饮食睡眠可，查体：神志清，精神可，双肺呼吸音稍粗，未闻及干湿性啰音。心律齐，未闻及杂音。腹部平坦，全腹软，无压痛、反压痛，双下肢无水肿。

出院诊断(Discharge Diagnosis):

1.2型糖尿病2.右肾囊肿3.甲状腺结节4.动脉粥样硬化5.脂肪肝6、高脂血症

Predictions

甲状腺结节 E04.101

2型糖尿病 E11.9022

高脂血症 E78.501

动脉粥样硬化 I70.904

脂肪肝 K76.001

肾结石 N20.000

后天性肾囊肿 N28.100

(a)

出院小结(Discharge Summary)

入院情况:（Admission Situation）

患者中年男性，因“体检发现血糖升高5年，血糖控制不佳6月”入院。患者5年前常规体检发现血糖升高，空腹约9mmol/L，当地诊断为:“2型糖尿病”。无明显口干、多饮、多尿及体重下降，无头晕、无尿泡沫增多、无手足麻木不适，偶有视物模糊。予以饮食及运动控制血糖，平素不规律监测血糖，空腹血糖波动于8-9mmol/L。病程中患者一般情况可，饮食睡眠可，大小便未见明显异常，近2月体重下降约3Kg。辅助检查B超：甲状腺右叶极低回声结节，考虑胶质囊肿；双侧颈动脉粥样硬化伴右侧斑块形成；脂肪肝；右肾囊肿伴钙乳；前列腺增大；左下肢动脉粥样硬化。

入院诊断:（Admission Diagnosis）

1.2型糖尿病2.肾结石3.甲状腺结节4.动脉粥样硬化5.脂肪肝

出院情况:（Discharge Situation）

患者无口干多饮，无多尿，饮食睡眠可，查体：神志清，精神可，双肺呼吸音稍粗，未闻及干湿性啰音。心律齐，未闻及杂音。腹部平坦，全腹软，无压痛、反压痛，双下肢无水肿。

出院诊断:（Discharge Diagnosis）

1.2型糖尿病2.右肾囊肿3.甲状腺结节4.动脉粥样硬化5.脂肪肝6、高脂血症

(b1)

出院小结(Discharge Summary)

入院情况:（Admission Situation）

患者中年男性，因“体检发现血糖升高5年，血糖控制不佳6月”入院。患者5年前常规体检发现血糖升高，空腹约9mmol/L，当地诊断为:“2型糖尿病”。无明显口干、多饮、多尿及体重下降，无头晕、无尿泡沫增多、无手足麻木不适，偶有视物模糊。予以饮食及运动控制血糖，平素不规律监测血糖，空腹血糖波动于8-9mmol/L。病程中患者一般情况可，饮食睡眠可，大小便未见明显异常，近2月体重下降约3Kg。辅助检查B超：甲状腺右叶极低回声结节，考虑胶质囊肿；双侧颈动脉粥样硬化伴右侧斑块形成；脂肪肝；右肾囊肿伴钙乳；前列腺增大；左下肢动脉粥样硬化。

入院诊断:（Admission Diagnosis）

1.2型糖尿病2.肾结石3.甲状腺结节4.动脉粥样硬化5.脂肪肝

出院情况:（Discharge Situation）

患者无口干多饮，无多尿，饮食睡眠可，查体：神志清，精神可，双肺呼吸音稍粗，未闻及干湿性啰音。心律齐，未闻及杂音。腹部平坦，全腹软，无压痛、反压痛，双下肢无水肿。

出院诊断:（Discharge Diagnosis）

1.2型糖尿病2.右肾囊肿3.甲状腺结节4.动脉粥样硬化5.脂肪肝6、高脂血症

(b2)

Figure 2: The screenshot of Clinical-Coder system, the English version can be found in the appendix A. (a) gives the predicted diseases after users enter the clinical notes which contains four parts, admission situation, admission diagnosis, discharge situation and discharge diagnosis. (b1) and (b2) are the visualization of supporting information for predictions.

words as implicit semantic features. The two kinds of semantic features are both clues to explain the reason why to assign each code, which is also the basis of experts in manual coding process. To capture the two semantic phenomena, we exploit dilated convolution and n-gram matching mechanism to extract implicit semantic features and explicit semantic features, respectively. Furthermore, we develop a system to assist the professional coders in assigning the correct codes. In summary, the main contributions are as follows:

- We collect a large-scale Chinese Clinical notes dataset, making up for the lack of Chinese ICD coding corpus.
- We propose a novel method to simultaneously capture implicit and explicit semantic features, which enables to give interpretability for each predicted code.
- We develop an open-access online system, called clinical-coder, that automatically assigns codes to the free-text clinical notes with an indication of the supporting information for each code to be predicted. It uses vivid visualization to provide interpretability of prediction for each ICD code. The site can be accessed by `http://159.226.21.226/disease-prediction`, and instructions video is provided at `https://youtu.be/U4TImTwEysE`.

Figure 2 illustrates an example of the automatic coding for a Chinese Clinical note in our system (For the convenience of readers, the English version is included in the appendix A). The left of Figure 2 (a) is the free-text notes user entered, and the right of Figure 2 (a) is predicted codes and corresponding disease names. Figure 2 (b1) and Figure 2 (b2) are the visualization of supporting information for predictions. The detailed description is presented in the section 3.2.

2 Related Work

2.1 Automatic ICD coding

Automatic ICD coding has recently been a research hotspot in the field of clinical medicine, where neural network architecture methods show promising results than traditional machine learning methods.

Most studies treat automatic ICD coding as a multi-label classification problem and use only the free-text in summaries to predict codes (Subotin and Davis, 2015; Kavuluru et al., 2015; Yu et al., 2019), while many methods benefit from extra information. Shi et al. (2017) encode label description with character-level and word-level long short-term memory network. Rios and Kavuluru (2018) encode label description with averaging words embedding. Furthermore, adversarial learning is employed to unify writing styles of diagnosis descriptions and ICD code descriptions (Xie et al., 2018). Besides code descriptions, Wikipedia comes to be regarded as an external knowledge source (Prakash et al., 2017; Bai and Vucetic, 2019).

Additionally, inferring interpretability is a crucial challenge and obstacle for practical automatic coding, since professionals are willing to be con-

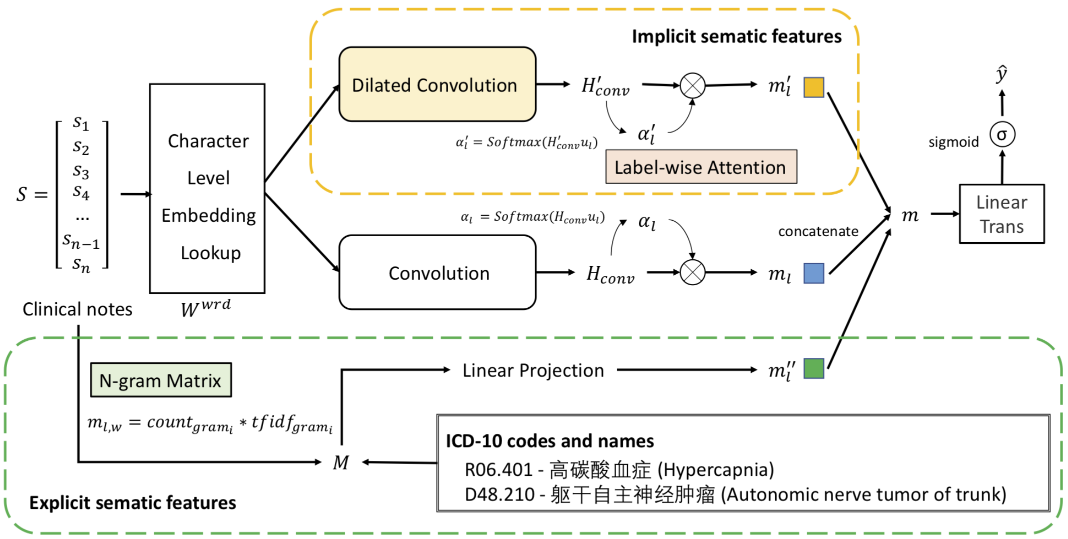

Implicit sematic features
Dilated Convolution
H'_{conv}
m'_l
$\alpha'_l = Softmax(H'_{conv}u_l)$
α'_l
Label-wise Attention
$\hat{y}$
sigmoid
σ
$S =$
s_1
s_2
s_3
s_4
...
s_{n-1}
s_n
Character Level Embedding Lookup
Clinical notes
W^{wrd}
$\alpha_l = Softmax(H_{conv}u_l)$
α_l
concatenate
m
Linear Trans
Convolution
H_{conv}
m_l
N-gram Matrix
Linear Projection
m''_l
$m_{l,w} = count_{gram_i} * tfidf_{gram_i}$
ICD-10 codes and names
R06.401 - 高碳酸血症 (Hypercapnia)
D48.210 - 躯干自主神经肿瘤 (Autonomic nerve tumor of trunk)
M
Explicit sematic features

Figure 3: The whole architecture of the model. The input is the clinical text, and output is the ICD codes. The yellow dotted box indicates how to use attention-based dilated convolution to capture the implicit semantic of non-continuous words. The green dotted box indicates how to use n-gram matching mechanism to capture the explicit semantic of continuous n-gram words.

vinced by the model insights of vital supporting information or decision-making process (Vani et al., 2017; Mullenbach et al., 2018). Baumel et al. (2018) employ bidirectional Gated Recurrent Unit with sentence-level attention to obtain relevant sentences for each code. Mullenbach et al. (2018) use attention at the word level, which is more fine-grained. Our work is inspired by (Mullenbach et al., 2018), assigning the importance value for each label to the discharge summaries to assists in explaining the model prediction process.

2.2 Dilated Convolution

Dilated convolution is designed for image classification to aggregate multi-scale contextual information without losing resolution in computer vision (Yu and Koltun, 2016). It inserts “holes” in the standard convolution map to increase the reception field. The hole-structure brings a breakthrough improvement to the semantic segmentation task.

Similarly, several hole-structured convolution neural networks (CNNs) (Lei et al., 2015; Guo et al., 2017) are designed to handle natural language processing tasks. In the text, there exists non-continuous semantic where useless information may be interspersed among the sentences. Holes in the dilated convolution can ignore the extra word between the non-continuous words and well adapt to match non-continuous semantic. Since the semantic infomation is crutial when understanding natural language(Zuo et al., 2019), we apply the dilated convolution to encode the text, capturing the non-continuous semantic information.

3 Clinical Coder System

3.1 Method

We propose a **D**ilated **C**onvolutional **A**ttention network with **N**-gram **M**atching Mechanism (DCANM) for ICD coding task. Figure 3 describes the architecture of the model. The input of model is all sentences in clinical notes, which are spliced together. The input sentences interact with ICD code names to capture explicit semantic features and generate an n-gram matrix. At the same time, the input sentences are transformed into vector and processed by dilated CNN to capture implicit semantic features. Attention mechanism is used to improve the performance. Then all features are concatenated to form the final features. Finally, we use a sigmoid classifier to predict the probability of each code. Next, we give the detailed descriptions.

Word Embedding. Word embedding is a low-dimensional vector representation of a word. We use the pre-trained embedded matrix $W^{wrd} \in \mathbb{R}^{d^w \times |V|}$, where d^w is the dimension of word embedding and $|V|$ is the size of vocabulary. Given a sentence, $S = [w_1, w_2, ..., w_N]$, where N is the number of words in the sentence, we can get the word embedding by:

$$w_e = W^{wrd} v_i, \quad (1)$$

where v_i is the one-hot representation of the current word in the corresponding column of W^{wrd}.

Explicit Semantic Features. N-gram matching mechanism is applied to capture explicit semantic features. We use disease names (D) to sampling on the text (T). First, move the sliding window on the disease name $d_l \in D$ to get a n-gram substring. Then, calculate the frequency of each n-gram substring in the free-text. The sum of frequencies of gram with same length n (denoted as $gram_n$) has reflected the emergence of disease names in the text, nevertheless some grams have their unique particularity. For example, given a 2-gram string, "糖尿" (Diabetes) is more representative than "慢性"(Chronic) though they have the same length. To represent the degree of importance of different n-gram, each n-gram is given a term frequency-inverse document frequency (tf-idf) weight. Finally, for each free-text clinical note, we calculate an explicit semantic n-gram matrix (M) with size of $L \times W$, where L is numbers of labels and W is the numbers of sliding windows. For example, we have four sliding windows which lengths are 2, 3, 4, 5, so W is 4. For the l-th row the w-th column item in the feature map, we have:

$$m_{l,w} = \sum_{i=1}^{L_{gram_{ln}}} count_{gram_{lni}} * tf\_idf_{gram_{lni}} \quad (2)$$

$$tf\_idf_{gram_i} = \frac{n}{L_{n_l}} * \frac{L}{freq_{gram_{lni}}}, \quad (3)$$

where w is the index of n-length sliding window, $gram_{ln}$ is all n-length substrings of the l-th disease name, $gram_{lni}$ is the i-th $gram_{ln}$, $L_{gram_{ln}}$ is the number of $gram_{ln}$, $count_{gram_{lni}}$ is the frequencies of $gram_{lni}$ in the text, L_{n_l} is the length of the l-th disease name, $freq_{gram_{lni}}$ is the frequencies of $gram_{lni}$ in all disease names.

In this calculation, we can distinguish the importance degree of n-gram substring. It also works on English clinical notes, for instance, in a specific case from MIMIC-III (Johnson et al., 2016), the tf-idf value of "history of" is 1.79 while "atrial fibrillation" is 9.32 because "history of" appears 249 times in all ICD disease names and "atrial fibrillation" only appears two times. The higher the value is, the more representative the word is. Therefore "atrial fibrillation" is more likely to indicate a disease than "history of".

Implicit Semantic Features. Dilated convolution is applied to capture implicit semantic features. For a long clinical text, dilated convolution extends the reception field in the situation of not using pooling operation so that every kernel has a wider range of information. More importantly, it has "holes" in convolution map, which means it can be adapted to match the non-continuous semantic information. For example, "类风湿性多部位关节炎"(Rheumatoid multisite arthritis) in the clinical notes refers to "类风湿性关节炎"(Rheumatoid arthritis) in ICD, the convolution map with holes can tolerate the redundant parts, as shown in Figure 4. It is a distinct advantage of dilated convolution for processing texts.

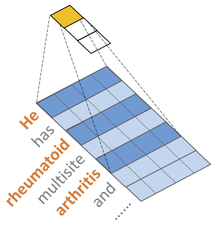


Figure 4: An example of the dilated convolution in processing text.

Formally, the actual filter width of dilated convolutional neural network is computed as,

$$k_d = r(k-1) + 1, \quad (4)$$

where $r \in [1, 2, 3, ...]$ is the dilated rate, k is the origin filter width.

For each step n, the typical convolution is computed as formula 5 and dilated convolution is computed as formula 6. The dilated CNN is same as typical CNN when the dilated rate is 1, since k_d equals to k when $r = 1$:

$$h_n = tanh(W_c * x_{n:n+k-1} + b_c) \quad (5)$$

$$h'_n = tanh(W_c * x_{n:n+k_d-1} + b_c), \quad (6)$$

where $W_c \in \mathbb{R}^{k_d \times d_e \times d_c}$ is the convolutional filter map, k_d is the actual filter width, d_e is the size of the word embedding, an d_c the size of the filter output and $b_c \in \mathbb{R}^{d_c}$ is the bias.

Attention. After convolution, the sentence is represented as $H \in \mathbb{R}^{d_c \times N}$. We employ the per-label attention mechanism (Mullenbach et al., 2018) to find the most contributed characters for each label.

For each label l, the distributed attention weight is computed as:

$$\alpha_l = SoftMax(H^T u_l), \quad (7)$$

where $u_l \in \mathbb{R}^{d_c}$ is the vector representation of label l. Finally, the sentence is represented as:

$$m_l = H\alpha_l \tag{8}$$

We employ attention both for typical CNN and dilated CNN, for convenience of distinction, we denote them as m_l and m'_l, respectively.

Classification. m_l and m'_l is concatenated with the linear transformed n-gram matrix horizontally. The aim of this step is to combining all the features together. Then we exploit sigmoid classifier and the prediction of label i is computed as,

$$\hat{y_i} = \sigma(W^T[m_l; m'_l; m''_l] + b), \tag{9}$$

where $i \in [1, 2, ..., L]$, $W \in \mathbb{R}^{3d_c}$, b is the bias, m''_l is the linear projection of n-gram matrix(M).

The loss function is the multi-label binary cross-entropy (Nam et al., 2013).

$$\mathcal{L} = \sum_{i=1}^{L}[-y_i log(\hat{y_i}) - (1 - y_i) log(1 - \hat{y_i})], \tag{10}$$

where $y_i \in \{0, 1\}$ is the ground truth for the i-th label and $\hat{y_i}$ is the sigmoid score for the i-th label.

3.2 User Interface

Figure 2 illustrates the user interface of our system.

User Input. The left of Figure 2(a) displays the user input. The user enters the whole free clinical note, which includes at least one from admission situation, admission diagnosis, discharge situation, and discharge diagnosis into the input box.

Predicted Labels. The predicted labels are presented in the list of Figure 2(a), including disease name and homologous ICD codes. The number of predicted codes are not always the same as the diseases in discharge diagnosis, because clinicians may leave out certain diseases and several diagnoses should be combined into one ICD code(Shi et al., 2017). Our model can list all these diseases, and give the reason why they should be predicted.

Interpretability. Interpretability is a critical aspect of the decision-making system, especially in the clinical medicine domain. In our system, we give two ways, n-gram matching mechanism and attention, to assist users in understanding why each code is predicted. A user can know why the model predicted the labels, and what the key information in its decision was:

(1) **N-gram Matching Mechanism.** When a patient suffering from a disease, the corresponding text span related to disease names often appear in the discharge summary. As shown in Figure 2 (b1), the gram in disease name is highlighted to give a hint to users if it appears in the clinical text. Highlighting not only tells users why we predict each code but also prompts the place of the important information.

(2) **Attention.** As shown in Figure 2 (b2), the red background is attention distribution, and the darker the color is, the more useful the word is to predict the current label. The darker color is also helpful and attractive for human-being to double-check the correction of labels.

4 Experiments

4.1 Dataset

We evaluate our model on both Chinese and English datasets. The Chinese dataset, collected by us, contains 50,678 Chinese clinical notes and 6,200 unique ICD-10 codes. For each clinical note, it contains five parts: admission situation, admission diagnosis, discharge situation, discharge diagnosis and annotated ICD-10 codes. Admission situation involves chief complaints, past medical history, etc. Discharge situation involves the results of general examination. Admission diagnosis and discharge diagnosis involve disease names, which may not be totally consistent with standard names in ICD-10. The manually annotated codes are based on ICD-10, which are tagged by professional coders after reading through the whole clinical note.

	CN-Full	CN-50	MIMIC-III-50
# Samples	50,678	36,758	9,795
# Labels	6200	50	50
Vocabulary size	3,957	3,957	51,917
# Average tokens per sample	621	655	1,530
# Average labels per sample	4.3	2.6	5.7

Table 1: Detailed information for three datasets.

The dataset (CN-Full) is formed with full labels mentioned above, and it is divided into train set and test set with the radio of 9:1. In addition, due to the phenomenon that massive codes are infrequent, and a small amount of codes are high-frequent, we reconstructed a sub-dataset (CN-50) with the most frequent 50 codes from the original dataset. The specific process is that filtering the origin train set and test set, and maintain the data which has at least one of the top 50 most frequent codes.

Dataset		CN-Full					CN-50					
Model	F1		AUC		R@k		F1		AUC		R@k	
	Macro	Micro	Macro	Micro	k=5	k=10	Macro	Micro	Macro	Micro	k=5	k=8
CAML(Mullenbach et al., 2018)	0.0600	0.6755	0.8832	0.9808	0.6099	0.7651	0.8305	0.8458	0.9846	0.9902	0.8796	0.9579
Dilated CNN	0.1017	0.6997	0.8637	0.9772	0.6268	0.7864	0.8399	0.8523	0.9849	0.9904	0.8807	0.9550
N-gram Matching	**0.1200**	0.7050	**0.9574**	**0.9915**	0.6393	0.8036	0.8385	0.8543	0.9867	0.9922	0.8900	0.9640
DACNM	0.1116	**0.7127**	0.9520	0.9909	**0.6430**	**0.8043**	**0.8452**	**0.8602**	**0.9878**	**0.9932**	**0.8895**	**0.9657**

Table 2: Evaluation on Chinese dataset CN-Full and CN-50.

To better compare with the previous works, we also evaluate our method on the MIMIC-III dataset (Johnson et al., 2016), which is the most authoritative English dataset for evaluating the performance of automatic ICD coding approaches. The detailed description for these datasets is listed in Table 1.

4.2 Data Preprocess and Parameters

We splice the admission situation, admission diagnosis, discharge situation and discharge diagnosis together, which is the input of the model. The max length of the input is 1000. The word embedding is pre-trained using Word2Vec (Mikolov et al., 2013) with the dimensions of 100. The text is from all clinical notes. The batch size is 16. The dropout rate is 0.5. The optimizer is Adam (Kingma and Ba, 2015) with a learning rate of 0.0001.

We use Micro-F1, Macro-F1, area under the ROC (Receiver Operating Characteristic) curve (AUC) and P@k as the metrics. P@k (Precision at k) is the fraction of the k highest-scored labels that are present in the ground truth.

4.3 Results

First, for the Chinese dataset (CN-Full and CN-50), CAML (Mullenbach et al., 2018) is set as our baseline, which use traditional convolutional attention network. Moreover, we test the dilated CNN and n-gram matching mechanism separately. The results in Table 2 indicate that dilated CNN and n-gram matching mechanism both have a positive effect on improving performance from baseline, and the best results are obtained when they combined.

We also evaluate our method on English dataset (MIMIC-III-50). The results are shown in Table 3. The CNN and Bi-GRU are the classic methods and the results are the same as (Mullenbach et al., 2018). Our proposed model achieves the Micro-F1 score of 0.641, which outperforms all previous works, more importantly providing interpretability.

Besides, we notice that macro-F1 measure is always lower than micro-F1, especially in the full labels datasets. It means the smaller classes have poorer performance than larger classes, which is consistent with the facts. Either MIMIC-III or the Chinese dataset, the sample distributions are extremely imbalanced. Minority of codes are highly frequent, while most codes are infrequent. N-gram matching mechanism helps improve macro-F1 on CN-Full dataset obviously, reaching two times than baseline. It can be inferred that utilizing grams in disease names is useful for the smaller class.

Model	F1		AUC		P@k
	Macro	Micro	Macro	Micro	k=5
CNN	0.576	0.625	0.876	0.907	**0.620**
Bi-GRU	0.484	0.549	0.828	0.868	0.591
C-MemNN(Prakash et al., 2017)	-	-	0.833	-	0.42
(Shi et al., 2017)	-	0.532	-	0.900	-
HA-GRU(Baumel et al., 2018)	-	0.366	-	-	-
CAML(Mullenbach et al., 2018)	0.532	0.614	0.875	0.909	0.609
DR-CAML(Mullenbach et al., 2018)	0.576	0.633	0.884	0.916	0.618
DACNM (Proposed model)	**0.579**	**0.641**	**0.890**	**0.916**	0.616

Table 3: Evaluation on MIMIC-III-50 dataset

5 Conclusion

In this paper, we propose a **D**ilated **C**onvolutional **A**ttention network with **N**-gram **M**atching Mechanism (DCANM) for automatic ICD coding. The dilated CNN, which is first applied to the ICD coding task, aims to capture semantic information for non-continuous words, and the n-gram matching mechanism aims to capture the continuous semantic. They both provide a pretty good interpretability for prediction. Moreover, we develop an open-access system to help users assign ICD codes. We will try to utilize external resources to solve the few-shot and zero-shot problem in the future.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61922085, No.61533018, No.61976211) and the Key Research Program of the Chinese Academy of Sciences (Grant NO.ZDBS-SSW-JSC006). This work is also supported by Beijing Academy of Artificial Intelligence (BAAI2019QN0301) and a grant from Ant Financial Services Group.

References

Tian Bai and Slobodan Vucetic. 2019. Improving medical code prediction from clinical text via incorporating online knowledge sources. In *The World Wide Web Conference*, WWW'19, 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 *Proceedings of the Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 409–416.

Jiahui Guo, Bin Yue, Guandong Xu, Zhenglu Yang, and Jin-Mao Wei. 2017. An enhanced convolutional neural network model for answer selection. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW'17, pages 789–790.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:16–35.

Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*, 65(2):155–166.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1511.07122*.

Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. 2015. Automatic icd-10 classification of cancers from free-text death certificates. *International Journal of Medical Informatics*, 84(11):956–965.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2015. Molding CNNs for text: Non-linear, non-consecutive convolutions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1565–1575.

Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*, pages 1–12.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Jinseok Nam, Jungi Kim, Iryna Gurevych, and Johannes Fürnkranz. 2013. Large-scale multi-label text classification - revisiting neural networks. In *Proceedings of the 2014 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, pages 437–452.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and No é mie Elhadad. 2013. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Aaditya Prakash, Siyuan Zhao, Sadid Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2017. Condensed memory networks for clinical diagnostic inferencing. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3274–3280.

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*, volume 2018, pages 3132–3142.

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

Michael Subotin and Anthony Davis. 2015. A method for modeling co-occurrence propensity of clinical codes with application to icd-10-pcs auto-coding. *Journal of the American Medical Informatics Association*, 23(5):866–871.

Ankit Vani, Yacine Jernite, and David Sontag. 2017. Grounded recurrent neural networks. *arXiv preprint arXiv:1705.08557*.

Pengtao Xie, Haoran Shi, Ming Zhang, and Eric P. 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.

Fisher Yu and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Ying Yu, Min Li, Liangliang Liu, Zhihui Fei, Fang-Xiang Wu, and Jianxin Wang. 2019. Automatic ICD code assignment of chinese clinical notes based on multilayer attention birnn. *Journal of Biomedical Informatics*, 91:103–114.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Event co-reference resolution via a multi-loss neural network without using argument information. *Science China Information Sciences*, 62.

A Appendix: English Version of Figure 2(a)

Since the online system is for Chinese clinical notes, we provide corresponding English version for reading convenience.

Admissions situation:

The middle-aged man was admitted to the hospital because of a “blood glucose increase of 5 years and poor glycemic control in June”. The patient’s routine physical examination five years ago revealed an increase in blood glucose, about 9 mmol / L on an empty stomach, and the local diagnosis was: “type 2 diabetes”. No obvious dry mouth, frequent drinking, polyuria and weight loss, no dizziness, no increased urine foam, no numbness in hands, feet, and occasionally blurred vision. Diet and exercise were used to control blood glucose. Normally, blood glucose was monitored irregularly, and fasting blood glucose fluctuated between 8-9mmol / L. During the course of the disease, the general condition of the patient is OK, the diet and sleep are OK, and there is no obvious abnormality in the stool. The weight loss in the last 2 months is about 3Kg. Auxiliary examination B: Ultralow hypoechoic nodules in the right lobe of the thyroid gland, considering glial cysts; bilateral carotid atherosclerosis with right plaque formation; fatty liver; right renal cyst with calcium milk; enlarged prostate; Atherosclerosis.

Admission diagnosis:

1. Type 2 diabetes **2.** Kidney stones **3.** Thyroid nodules **4.** Atherosclerosis **5.** Fatty liver

Discharge situation:

The patient had no dry mouth and frequent drinking, no polyuria, diet and sleep were OK, physical examination: clear mind, good spirits, slightly thicker breathing sounds in both lungs, and no wet and dry rales. Heart rhythm is uniform, and no noise is heard. The abdomen is flat, the whole abdomen is soft, no tenderness, no tenderness, no edema in both lower limbs.

Discharge diagnosis:

1.Type 2 diabetes **2.** Right renal cyst **3.** Thyroid nodule **4.** Atherosclerosis **5.** Fatty liver **6.** Hyperlipidemia

Predicted diseases and codes:

1. Thyroid nodule E04.101
2. Type 2 diabetes E11.9022
3. Hyperlipidemia E78.501
4. Atherosclerosis I70.904
5. Fatty liver K76.001
6. Kidney stones N20.000
7. Acquired renal cysts N28.100