Attentive Dual Embedding for Understanding Medical Concept in Electronic Health Record

Xueping Peng*, Guodong Long*, Shirui Pan†, Jing Jiang*, Zhendong Niu‡
* Centre for Artificial Intelligence, FEIT, University of Technology Sydney, Australia
† Faculty of Information Technology, Monash University, Australia
‡ School of Computer Science and Technology, Beijing Institute of Technology, China
Email: xueping.peng, guodong.long@uts.edu.au, shirui.pan@monash.edu, jing.jiang@uts.edu.au, zniu@bit.edu.cn

***Abstract*—Medical concept embedding is to learn a distributed representation for a medical related entity, e.g. diagnosis, treatment procedure, and medicine, which is a code stored in Electronic Health Record (EHR). The distributed representation is expected to preserve the comprehensive relationships among medical concepts rather than one-hot encoding, and it will be the inputs of machine learning based healthcare analytic tasks. Therefore, the performance of the analytic tasks highly depends on the quality of embedding outputs. To fully utilise the information in EHR, this paper proposes a novel attentive dual embedding method, namely MC2Vec, to intensively capture the proximity relationships among medical concepts. In particular, the proposed MC2Vec method uses a two-step optimisation framework to recursively refine the embedding via two components 1) Skip-gram based method to generate the initial embedding of medical concept, and 2) Attentive CBOW based method to fine-tune the code embedding by adding the temporal information of one patient's sequential healthcare activities. The experiment studies on two public EHR datasets demonstrate the effectiveness of the proposed MC2Vec method, which performs superior than other five state-of-the-art embedding methods.**

***Index Terms*—Medical Concept Embedding, Attention Mechanism, Med2Vec, Dual Embedding**

I. Introduction

The healthcare information system stores Electronic Healthcare Records (EHR) data to record patients' sequential healthcare visits, where each visit is a set of medical entities and concepts [2]. To standardise the healthcare procedure, these medical concepts are used to be converted to codes by using a standard coding system, such as diagnosis code, procedure code of treatment, and drug code in pharmacy. However, existing medical coding system is used to be a tree hierarchy defined by medical expert experience, and the tree structure is straightforward for human understanding and maintenance. This tree-based coding system includes the basic taxonomy knowledge of medical, but it discards the complex relationship among each unit of the medical concepts. In EHR dataset, there are many complicated co-occurrence relationship among medical concepts that includes much richer information than tree-based taxonomy. Therefore, the code derived from EHR data will provide more information for further healthcare analytics, for example, diagnoses prediction [1], [22], [25], predicting inpatient mortality [9] and length of stay after admission [9].

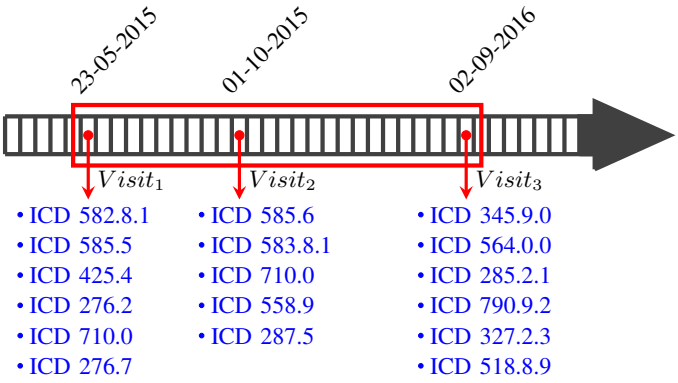

Fig. 1. An example segment of one patient's healthcare journey

The EHR data is used to be a multi-level structure including three layers: patient, visit and medical concept. A patient's healthcare journey, namely patient journey, is a sequence of visits occurring in different time stamp. Each visit record is composed of a set of medical concepts or codes. The Fig. 1 shows an example segment of one patient journey EHR [5], [25] in EHRs, and each visit has a set of medical concepts, e.g. International Classification of Diseases (ICD). In Natural Language Processinig (NLP), there is a similar multi-level structure with document, sentence and word. In particular, a document is composed of a sequence of sentences, and each sentence is a bag of words. However, there is some summarised differences between two domains.

- The visits in one patient journey are sequential with time interval, and the sentences in one document just have sequential relationship.
- In the bag of medical concepts, there is one dominated ICD code that is the principle disease, and other codes share equal importance and have no sequential relationship. The words in a sentence have sequential relationship and all of them share the same importance weight.
- Each medical code in a bag is unique, and the sentence may includes repeat words.

To effectively utilise the semantic information of EHRs, a medical concept embedding method is desired to develop machine learning based medical applications. The one-hot encoding of medical concept will generate a vector with

high-dimension and sparsity. A straightfoward solution is to use the word embedding approaches to learn representations of medical concepts [7], [12], [14], [15] and it has been used to improve the performance of various healthcare applications [16], [17], [19], [20], [22]. Choi Edward, et al [1], [31] proposed a multi-level representation learning to simultaneously embedding the visits and medical concepts by using the sequential order of visits and co-occurrence of medical concepts. X. Cai et al [5] proposed a CBOW based medical concepts embedding method enhanced with attention mechanism to capture the temporal information of visits. In particular, the temporal sequence of patient visits has been split into many time units so that the attention mechanism can capture the sequential information as well as the time-aware information. However, a fixed size of time units is impractical because different diagnosis or treatment might have different awareness of time. Moreover, a large size of time units may cause information loss because it puts several visits into one time unit, and a small size of time units will cause dimension explosion on attention mechanism.

To fully utilise the information in multi-level EHRs, we propose a novel attentive dual embedding to embed medical concept to vector, namely **MC2Vec** (Medical Concept to Vector). This dual embedding model is controlled by an novel loss function that is designed to satisfy three objectives: 1) using target medical concept to accurately predict the context of the concept, namely One-to-N embedding, 2) using the context of the medical concepts to accurately predict the target concept, namely N-to-One embedding, and 3) considering the temporal sequential information by using attention mechanism. In particular, the attentive dual embedding method uses a two-step optimisation to discover the optimal solution, the embedding result, by two steps. Firstly, we convert the medical concept to a representation, then apply Skip-gram to embed the medical concept by using one concept to predict context concepts, and generate embedding results as One-to-N embedding. Then, we use the One-to-N embedding to represent the medical concept, and train the Attentive CBOW model to fine-tune the embedding as N-to-One embedding. The above two steps will be conducted recursively in the overall framework, and the initial value of medical concepts' representation is one-hot encoder.

The paper is an attempt to tackle medical concept embedding task, and its contributions are summarized as below.

- A novel dual embedding method that can fully utilise the information in EHRs, and a new loss function is proposed to optimise the pipeline procedure of two embedding models.
- A proposed attentive CBOW method to capture the temporal information in a flexible way and with less information loss on time interval.
- A new practical medical concept method that achieves the-start-of-the-art performance in two public datasets.

The remainders are organized as follows. In Section II, we briefly discuss some related work. Then, details about our method are presented in Section III. In Section IV, we demonstrate the experimental results conducted on real world public datasets. Finally, we conclude our study and prospect our future work in Section V.

II. Related Work

A. Word Embedding

Although word embedding was first introduced by Rumelhart et al. [4] in 1986, distributed representation learning of words with neural network based models has became a hot research topic since 2003 [3], [7], [12]–[15]. CBOW and Skip-gram model [12], [13] are two of the model families, which were introduced to compute continuous vector representations of words from very large data sets. These two models have an assumption that the order of words in the context does not influence the projection of the target word. Recently, some research has studied the influence of word context in neural networks. For example, Melamud et al. [21] and Liu et al. [23] explored the impacts of context types and the target word conditioned on a subset of the contexts for the Skip-gram model, respectively. Ling et al. [18] extended CBOW by integrating attention model to consider contextual words and its relative position to the predicted word. As discussed in Section I, there are 3 significant differences between document and patient journey. Hence, it would cause information loss if we directly apply word embedding models to learn representations of medical concepts.

B. Medical Concept Embedding

Borrowing the ideas from word representation models [12], [13], researchers have recently explored the possibility of efficient representations for medical concepts in the healthcare domain. Skip-gram model has been directly exploited to learn the representations of medical text [16] and UMLS medical concepts [17]. Choi et al. [20] applied the Skip-gram model to learn medical concepts embeddings from different data sources, such as medical journals, medical claims and clinical narratives. Choi et al. [1] introduced a med2vec model based on the Skip-gram to learn concept-level and visit-level representations simultaneously. All these models took EHRs as documents without considering the temporal information. Recently, The attention mechanism [8] has been introduced to healthcare analytics. A graph-based attention model [22] has been proposed to incorporate medical ontologies to learn representations of medical concepts; Rajkomar et al. [9] applied an attention-based time-aware neural network model to predict patient outcomes; Cai et al. [5] proposed MCE to integrate time information into the attention model to embed medical concepts. Our work differentiates from the previous since time-aware attention we focus on time gap between visits of EHRs, and the context window is not based on time units but temporal window.

III. Proposed Model

This section starts by introducing some definitions of medical concepts and the related notations, then briefly introduce

the basic units for medical concept embedding. The final part is to describe the proposed attentive dual embedding method.

A. Preliminary

Definition 1 (Medical Concept): A medical concept is defined as a term or code to describe diagnosis, procedure, medication, and laboratory tests for an inpatient during a treatment process. We denote the set of medical concepts as $C = \{c_1, c_2, ..., c_N\}$, where *N* is the size of medical concepts in the dataset.

Definition 2 (Visit): A visit for an inpatient refers to a treatment process from admission to discharge, including an admission time stamp. We denote a visit as $V_t = \{c_{t,1}, c_{t,2}, ..., c_{t,K}\}$, where $c_{t,i} \in C, i = 1, ..., K$, *K* is the size of medical concepts in a visit and *t* is admission time.

Definition 3 (Patient Journey): A patient journey consists of a sequence of visits over time, which is denoted as $J = \{V_{t_1}, V_{t_2}, ..., V_{t_M}\}$, where *M* is the total visit times for a patient .

Definition 4 (Temporal Interval): Temporal interval refers to time difference between two visits in a patient journey, which is denoted as $\triangle = |t_i - t_j|$, where $i, j = 1, ..., M$.

Definition 5 (Task): Given a set of Patient Journey *Js*, the task is to learn an embedding function $f_C : C \rightarrow R^d$ that maps every code in the set of medical concept *C* to a real-valued dense vector with dimension *d*.

B. Basic Units for Medical Concept Embedding

The most straightforward embedding method is to adapt the classic embedding models [12], [13], CBOW and Skip-gram, to tackle medical concept embedding task. The fundemental idea is to generate training samples from EHRs by select one medical concept as target vector, and its co-occurrence or co-related medical concepts as context. The CBOW-based medical concept embedding method is to learn the representations by constructing a neural-based classification model that uses the context vector including mutliple medical concepts to predict the target word, also named as N-to-One embedding. The other is Skip-gram, instead of predicting the target word based on the context, uses each target vector as an input to predict context vector, also named as One-to-N embedding.

Given EHRs's multi-level structure, we extract training samples by using the sequential visits and co-occurence of medical concepts. Each visit V_t is a bag of medical concepts $\{c_1, c_2, ...\}$. Each c_i in the bag will be a target vector and its context vector $H = \{c_k, c_l, c_m, c_n, ...\}$ will be randomly sampled medical concepts from this bag. Sometimes, we will give the first medical concept a bigger probability of sampling because the first one is used to be the dominate item of the bag, e.g. the first diagnosis code is the majority disease dominate the visit. We also can use a slide window to select the related medical concepts based on the assumption that the medical doctors are used to write the highly co-related concepts together. The choosing of these alternative sampling methods is based on the empirical analysis of the dataset. To facilitate the description, this paper choose the slide window as the sampling method.

1) CBOW-based medical concept embedding: The objective of CBOW is to maximize the average log probability of the occurrences of target vector w given context vector H. With a given visit, we can define an objective function as maximal likelihood estimation.

$$\max \frac{1}{T-2k} \sum_{t=k}^{T-k} \log p(c_t|H_t), \tag{1}$$

where T is the total number of medical concepts in the given visit, k is the size of slide window, and H_t is the context vector that select the medical concepts by using slide window.

2) Skip-gram-based medical concept embedding: The objective of the Skip-gram model is to maximize the average log probability of predicting context vector by using the target vector, and its objective function can be defined as below.

$$\max \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{K} \log p(c'_j|c_t), \tag{2}$$

where K is the total number of medical concepts in context vector, c'_j is a medical concept in context vector.

The probability $p(c_t|H_t)$ in Equation 1 and $p(c'_t|c_t)$ in Equation 2 can be defined as the Softmax function. Regardless the CBOW-based N-to-One embedding and the Skip-gram-based One-to-N embedding, it can be generalized to the definition: use an input vector w_I to predict w_O, and the probability of prediction is a Softmax function as below.

$$p(w_O|w_I) = \frac{\exp\{v'^{T}_{w_O} v_{w_I}\}}{\sum_{w=1}^{W} \exp\{v'^{T}_{w} v_{w_I}\}}, \tag{3}$$

where v_w and v'_w are the "input" and "output" vector representations of w, and *W* is size of the medical concept vocabulary. For CBOW, $v_{w_I} = (1/2c) * \sum_{w_j \in H_n} v_{w_j}$.

3) Negative Sampling: The formulation 3 is impractical because the cost of computing $\bigtriangledown \log p(w_O|w_I)$ is proportional to *W*, which is often large. To reduce the computational complexity, the word2vec model uses negative sampling to replace every $\log p(w_O|w_I)$ term in CBOW and Skip-gram objectives, which instead maximizes:

$$J = \log \sigma(v'^{T}_{w_O} v_{w_I}) + \sum_{i=1}^{r} \mathbb{E}_{w_i \sim P(w)} [\log \sigma(-v'^{T}_{w_i} v_{w_I})], \tag{4}$$

where σ is the Sigmoid function, r is the number of negative samples, and $P(w)$ is the noise distribution [12].

C. Attentive Dual Embedding Approach

Different medical concepts in a patient journey have temporal relationships that is an important information for embedding. The *One-to-N* and *N-to-One* embedding results can capture different view of the comprehensive relationship. Therefore, we propose an attentive dual embedding method that can capture multiple view of semantic relationship and grasp the temporal information.

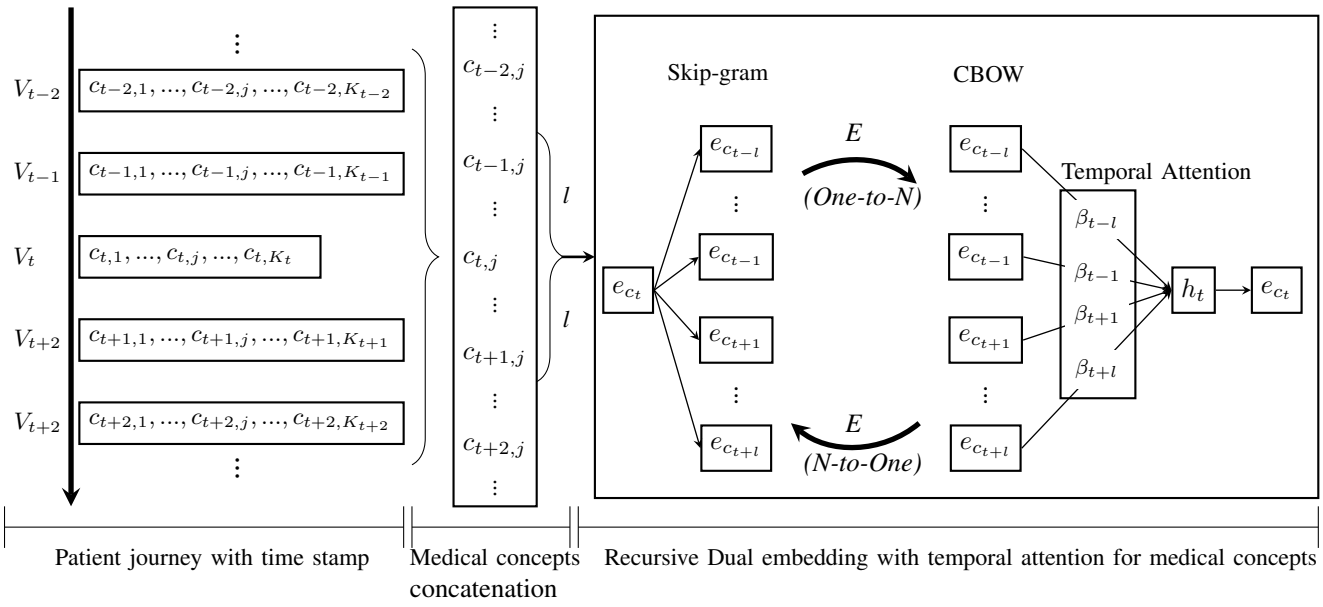

Fig. 2. Dual Embedding model for medical concept with temporal attention. There are 3 stages: Stage 1 to decompose patient journey into time sequential visits; Stage 2 to concatenate concepts in each visit in a patient journey as patient vector; Stage 3 to conduct dual embedding of medical concepts on a temporal skip window *l*, integrating three components of Skip-gram, CBOW and temporal Attention, where *E* is the embedding parameters.

1) Approach Architecture: The framework's architecture is shown in Fig. 2. The attentive dual Embedding model for medical concept, also named as MC2Vec, has three parts including a) Patient journey with time stamp, b) medical concepts concatenation, and c) Attentive dual embedding.

a) Patient journey with time stamp: We split the patient journey into *M* visits, i.e. $V_t = \{c_{t,1}, ..., c_{t,j}, ..., c_{t,K_t}\}$, where *t* is patient visiting time to hospital. Each medical concept *c* is associated with a time stamp *t*.

b) Medical concepts concatenation: To generate a context and target concept for MC2vec, visits of a patient journey are concatenated according to its' temporal sequence to a vector of medical concepts with time stamp. For example, a patient has three visits, $V_1 = \{c_{t_1,1}, ..., c_{t_1,j}, ..., c_{t_1,K_{t_1}}\}$,$V_2 = \{c_{t_2,1}, ..., c_{t_2,j}, ..., c_{t_2,K_{t_2}}\}$,$V_3 = \{c_{t_3,1}, ..., c_{t_3,j}, ..., c_{t_3,K_{t_3}}\}$. The concatenated vector would be $J_{vec} = \{c_{t_1,1}, ..., c_{t_1,j}, ..., c_{t_1,K_{t_1}}, c_{t_2,1}, ..., c_{t_2,j}, ..., c_{t_2,K_{t_2}}, c_{t_3,1}, ..., c_{t_3,j}, ..., c_{t_3,K_{t_3}}\}$.

c) Dual embedding for medical concepts: Given J_{vec}, temporal window size *l* and target concept c_t, we firstly exploit Skip-gram with a temporal skip window to learn the embedding parameters *E* of medical concepts over the context as *One-to-N* embedding, then employ the *One-to-N* embedding *E* and the temporal attention to learn the representations for medical concepts with attentive CBOW in the same skip window, denoted as *N-to-One*. The *One-to-N* works like expectation step in EM algorithm [6], which fixes embedding parameter of target concept c_t to optimize the embedding parameters of its context concepts. And the *N-to-One* like the maximization step of EM algorithm, embedding parameters of target concept c_t are optimized by fixing embedding parameters of the context concepts. By sliding the temporal window *l* over J_{vec} to get different target concept c_t, the *One-to-N* and *N-to-One* mutually reinforce each other to learn optimized embeddings for medical concepts.

Dual embedding consists of three components: Skip-gram, CBOW and temporal attention. On one hand, Skip-gram is better for infrequent medical concepts than CBOW [12], [13], on the other hand, attentive CBOW integrates temporal information to learn non-uniform attention weights within a temporal context. Therefore, our model can improve medical concept embeddings by identifying infrequent concepts and capturing temporal distributional relationships.

d) Unified training: To obtain optimized medical concept representation, the single unified framework can be obtained by adding the objective function of *One2N* (Skip-gram) and the one of *N2One* (attentive CBOW) as follows,

$$\max_{E} J_{One2N} + J_{N2One} \tag{5}$$

$$J_{One2N} = \sum_{c_j \in H_t} \{\log \sigma(e_{c_j}^{'T} e_{c_t}) + \sum_{i=1}^{r} \mathbb{E}_{c_i \sim P(c)}[\log \sigma(-e_{c_i}^{'T} e_{c_t})]\}$$

$$J_{N2One} = \log \sigma(e_{c_t}^{'T} h_t) + \sum_{i=1}^{r} \mathbb{E}_{c_i \sim P(c)}[\log \sigma(-e_{c_i}^{'T} h_t)]$$

where *E* is the embedding parameters, c_t the target medical concept and h_t the weighted context of c_t, c_x the negative sample, and $H_t = \{e_{c_{t-l}}, ..., e_{c_{t-1}}, e_{c_{t+1}}, ..., e_{c_{t-l}}\}$. By combining the two objective functions, we learn medical concept embeddings from the same temporal skip window.

2) Temporal Attention: To capture the temporal semantic relationships among medical concepts, we develop a temporal attention mechanism that can learn the non-uniform attention weights in a temporal skip window. Particularly, the prior Skip-gram model's embedding results will be the input of the attentive-based CBOW embedding model. The context vector is calculated by non-uniform weighting the context vectors:

$$h_t = \log(2l+1)\log(\sum_{e_{c_i}\in H_t}\beta_i^2)\sum_{e_{c_i}\in H_i}\beta_i e_{c_i} \tag{6}$$

where l is the temporal skip window, $\log(2l+1)$ and $\log(\sum_{e_{c_i}\in H_t}\beta_i^2)$ are scalars to the weighted sum $\sum_{e_{c_i}\in H_i}\beta_i e_{c_i}$.

$$\beta_i = \frac{e^{\alpha_i}}{\sum_{e_{c_j}\in H_t} e^{\alpha_j}} \tag{7}$$

For the attribution logits, we introduce k functions, $A_1(\triangle), ..., A_k(\triangle)$, where each A_i has the form $A(\triangle) = \log(\triangle + 1\text{day})$, and $\triangle$ is the temporal interval between each context $e_{c_i} \in H_t$ and the target e_{c_t}.

We define a k dimensional projection of the embedding by learning a $k \times d$ dimensional matrix P, and for $e_{c_j} \in H_t$ multiplying it to get the k scalars $p_{1,j}, ..., p_{k,j}$. We then define the attribution logits to be

$$\alpha_i = \sum_{i=1}^{k} p_{i,j} A_i(\triangle_j) \tag{8}$$

The model learns to pay more attentions on temporal interval, which can improve medical concept representations by identifying the related visit time interval and capturing more accurate related target-context pairs.

3) Model Parameters and Complexity: We use Adam [27], one of gradient descent optimizers [27]–[30], to train our model. The parameters of this optimizer is default as the same for recommendation of Adam. Compared to the Skip-gram and CBOW model, the additional computation is temporal attentions. Algorithm 1 shows the details of our model. Note that each operation of computing an attention weight is to multiply P with e_{c_j}. Hence, the complexity of computing the attentions is related to the temporal attention window k, which will be discussed in Section IV.

Algorithm 1 Algorithm of MC2Vec Model

Input: Set of Patient Journey *Js*
Output: Medical Concept Embedding Parameters $E \subset \mathbb{R}^{N\times d}$

Initialization: $E^{(0)}$
for each $J \in Js$ **do**
  Initialization: J_{vec}
  for each $V \in J$ **do**
    push V into J_{vec}
  end for
  generate a batch of samples d from J_{vec}
  for i = 0 to $(|d|-1)$ **do**
    $E^{(2i+1)} = F(E^{(2i)})$ //F: Skip-gram function
    $E^{(2i+2)} = G(E^{(2i+1)})$ //G: Att. CBOW function
  end for
end for
return E

IV. Experiments

The performance of the proposed model will be evaluated on two public datasets via clustering task from machine learning.

A. Dataset Description

We conduct comparative studies on two public datasets listed as follows:

a) CMS: is a publicly available[1] synthetic claims dataset, which includes four types of files, such as inpatient, outpatient, carrier and the beneficiary summary. In the experiment, we only choose a sub-dataset of inpatient files between 2008 and 2010 as one of our two datasets. The basic statistical information is shown in Tab. I.

b) MIMIC III: is an open-source, large-scale, de-identified and ICU patients related EHR data set. The MIMIC III [24] dataset mainly consists of clinical logs of patients admitted to critical care units with serious conditions. The diagnosis codes in this dataset follow the ICD9[2] standard. The statistics of the dataset are provided in Tab. I.

TABLE I
Statistics of Datasets.

Datasets	CMS(08-10)	MIMIC III
# of patients	755,214	46,520
# of visits	1,332,822	58,976
Avg. # of visits per patient	1.76	1.27
# of unique diagnose codes	7,873	6,985
# of unique procedure codes	10,726	2,032

B. Ground Truth

The clustering task will be conducted to evaluate the quality of the embedding results. We choose ground truth by using two well-organized ontologies including ICD9 standard and Clinical Classifications Software (CCS) [3]. The ICD9 standard has a hierarchical structure [26] shown in Fig. 3. In particular, the first three numbers of all codes ranging from 460 to 519 are labelled as *diseases of the respiratory system*, which is one of 19 categories. We use the high level nodes as the clustering labels. We obtain 19 categories for MIMIC III and CMS dataset. This kind of ground truth is named as **ICD**.

CCS provides a way to classify diagnoses and procedures into a limited number of categories by aggregating individual

[1]https://www.cms.gov
[2]http://www.icd9data.com
[3]https://www.hcup-us.ahrq.gov

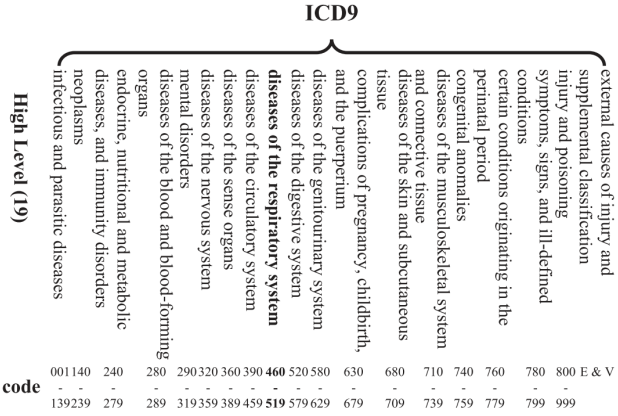

Fig. 3. The hierachical structure of ICD9.

ICD9 codes into broad diagnosis and procedure groups to facilitate statistical analysis and reporting [4]. CCS aggregates ICD9 diagnosis codes into 285 mutually exclusive categories. Examples of CCS diagnosis categories are shown in Tab. II. We obtain 265 categories for MIMIC III and 274 for CMS, respectively. We refer to this set of ground truth as **CCS**.

TABLE II
EXAMPLES OF CCS DIAGNOSIS CATEGORIES

Description	ICD9 Diagnosis Codes	CCS Category
Essential Hypertension	4011 4019	98
Hypertension with complications and secondary hypertension	4010 40200 40201 40210 40211 40290 40291 4030 40300 40301 4031 40310 40311 4039 40390 40391 4040 40400 40401 40402 40403 4041 40410 40411 40412 40413 4049 40490 40491 40492 40493 40501 40509 40511 40519 40591 40599 4372	99

C. Baseline Methods

We compare the proposed model with another 5 baseline models that are state-of-the-art embedding methods as below list, and all baseline models have been trained with their source codes.

a) ***CBOW-based medical concept embedding*** *(CBOW):* To learn the representations by averaging the context within a sliding window to predict the target vector.

b) ***Skip-gram-based medical concept embedding*** *(Sg):* To predict the target vector based on the context, which uses each target word as an input to predict words within context.

c) ***GloVe*** *[1]:* An unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

[4]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf

d) ***med2vec*** *[1]:* A multi-level embedding model to embedding medical concepts and visits simultaneously.

e) ***MCE*** *[5]:* A CBOW model with time-aware attention model to embed medical concepts with temporal information.

f) ***CBOW_Attn****:* A model based on CBOW to integrate sequential visit temporal interval into the attention model to learn representations of medical concepts, which is attentive CBOW in our proposed model.

g) ***Sg_CBOW****:* Our proposed vanilla dual embedding model for medical concepts with Skip-gram and CBOW without attention mechanism.

h) ***MC2Vec****:* Our proposed attentive dual embedding model for medical concepts with Skip-gram and attentive CBOW to learn representations of medical concepts.

All infrequent medical concepts will be removed and the threshold is emprically set to 5. Following the original Word2vec [12], [13], the same negative sampling strategy is used for Skip-gram and CBOW, CBOW_Attn, Sg_CBOW and MC2Vec, and the number of negative samples of MIMIC III and CMS is set to 10 and 5 respectively. All models are trained with 10 epochs for MIMIC and 5 epochs for CMS. The dimension d of medical concept embedding is set to 100.

D. Results

We use the clustering task to evaluate the embedding results on two public datasets, MIMIC III and CMS. We choose K-Means as the clustering algorithm, and use clustering performance indicator to evaluate the learned representations for medical concept. The temporal skip window of our model is emprically set to 9 for both datasets. We use the two sets of ground truth to evaluate the embedding performance of the proposed model and other baselines.

TABLE III
CLUSTERING PERFORMANCE (NMI) OF THE MODELS ON TWO DATASETS W.R.T. GROUND TRUTH ICD AND CCS (%).

Model	MIMIC III		CMS	
	ICD	CCS	ICD	CCS
CBOW	16.42	51.38	7.65	41.69
Sg	18.93	51.85	5.56	34.48
GloVe	19.18	48.24	7.58	34.11
med2vec	5.25	33.65	3.69	17.66
MCE	8.49	39.23	4.29	31.75
CBOW_Attn	23.20	54.77	12.48	43.82
Sg_CBOW	29.20	57.73	11.57	42.93
MC2Vec	**30.79**	**58.85**	**15.09**	**44.49**

a) Overall Performance: Normalized mutual information (NMI) for clustering performance is reported in Table III, where we highlight the best results. From Table III, it is concluded that the MC2Vec model obtains the best performances comparing with most state-of-the-art models on medical concept representation when the skip window is set to 9. The performance of MC2Vec outperforming the other models can be explained by introducing dual embedding model and incorporating the temporal attention in the model, which learns better embeddings of medical concept. Furthermore, it is noted that Sg_CBOW and CBOW_Attn are still competitive to

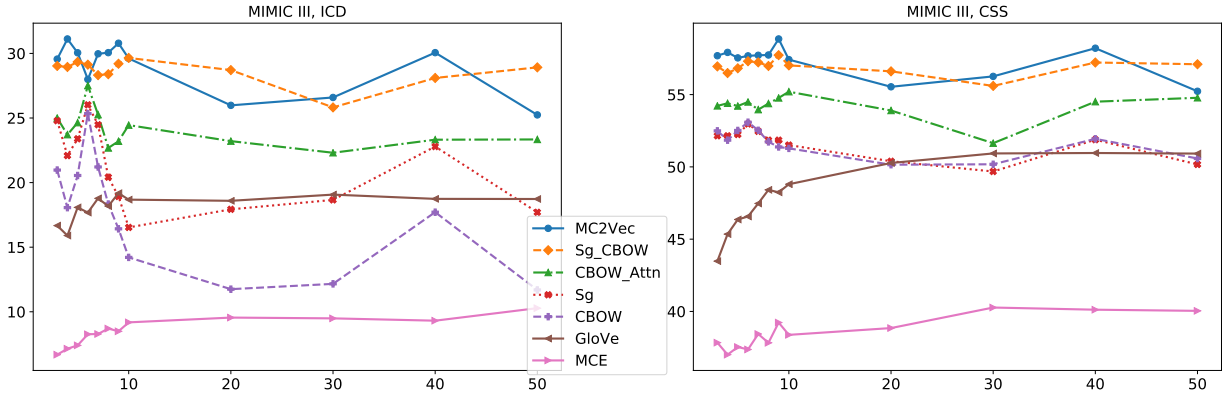

Fig. 4. NMI (%) of the models on MIMIC III w.r.t. ground truth ICD and CCS. The window size varies from 3 to 50.

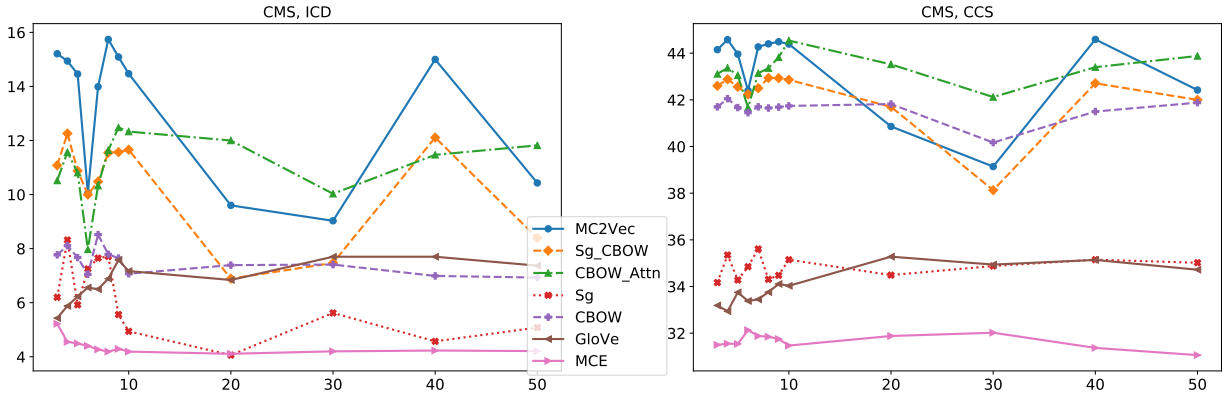

Fig. 5. NMI (%) of the models on CMS w.r.t. ground truth ICD and CCS. The window size varies from 3 to 50.

MC2Vec, because Sg_CBOW is our vanilla dual embedding model and CBOW_Attn is the *Output* part of dual embedding with temporal attention. Moreover, all models achieve better performance on the ground truth of CCS than that on ICD. It might be explained by that CCS has a well-organized ontology containing the experts' knowledge.

b) Performance of varying skip window sizes.: Considering effects of the context window to the performance of these models, we vary the context window size to compare the obtained performances. In this work, we only compare the proposed model MC2Vec with other six baselines, excepting med2vec due to lack of parameter for window size. Particularly, the window size is adjusted from 3 to 50 for each model.

The results on the clustering is summarized on MIMIC III dataset in Fig. 4. The performance of most models is decreased as the window size increasing that will induced noise. Because Glove takes use of global co-occurrences and MCE obtains bigger temporal scope, both of them are not sensitive on increasing window size. Moreover, the MC2Vec model and Sg_CBOW are competitive and always outperform the rest models in terms of NMI, which demonstrates that the integration of two embedding models can capture more comprehensive relationships among medical concepts. Specifically, as we increase the window size, Glove and MCE achieve better performances with larger window size.

Fig. 5 is the summary of results on the clustering task over CMS dataset. The MC2Vec model outperforms the baseline models in terms of NMI on the CCS ground truth when the skip window size is not more than 10, which demonstrates that the attention mechanism bring benefits to the the embedding in a smaller window. Particularly, the GloVe, Skip-gram, and MCE are relatively stable for changing window size. Other models obtain the local minimum values at skip window setting to 6, and achieve the best performances with window size setting to 8.

c) Influence of the Attention Window k: In this paper, temporal attention window is introduced to learn attention values from the time interval between visits. Fig. 6 shows the changes of MC2Vec's performance on different attention window size k that varies from 10 to 500. Two vertical axes are used to represent the different ranges of the results.

Fig. 6 shows that both ICD and CCS obtain the highest NMI values on MIMIC when the attention window is 300, and the performance drops quickly when the attention window greater than 300. It is due to the data sparsity in EHR data between 2001 and 2012. For CMS, the MC2Vec achieves the highest NMI on both ICD and CCS when the attention window is 100. It demonstrates the proposed method's effectiveness on

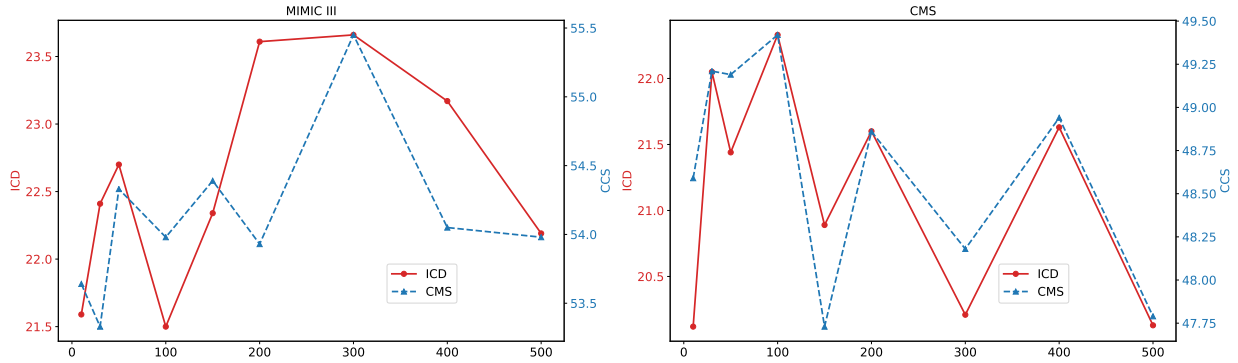

Fig. 6. NMI (%) on MIMIC III and CMS by varying attention windown k from 10 to 500.

large-scale and dense datasets.

V. Conclusion

This paper proposes an attentive dual embedding method, **MC2Vec**, that use dual embedding to capture the multiple view of the comprehensive relationships among medical concept, and use tailored attention mechanism to grasp the temporal information. The experimental study demonstrates the effectiveness of the proposed embedding method over two public datasets by comparing to baseline methods.

References

[1] E. Choi, M.T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, J. Sun, "Multi-layer representation learning for medical concepts," SIGKDD 2016: 1495-1504.

[2] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE journal of biomedical and health informatics, **22**(5), 1589-1604, 2018.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," ACL 2017 **5**: 135-146.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, **323**(6088), 5331, 1986.

[5] X. Cai, J. Gao, K. Y. Ngiam, B. C. Ooi, Y. Zhang, X. Yuan, "Medical Concept Embedding with Time-Aware Attention" IJCAI 2018: 3984-3990.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the royal statistical society. Series B (methodological), 1-38, 1977.

[7] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," JMLR, **3**(2), 1137-1155, 2003.

[8] D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate." arXiv:1409.0473, 2014.

[9] A. Rajkomar, et al., "Scalable and accurate deep learning with electronic health records," npj Digital Medicine **1**(1), 18, 2018.

[10] J. Sun, F. Wang, J. Hu, S. Edabollahi, "Supervised patient similarity measure of heterogeneous patient records," ACM SIGKDD Explorations Newsletter, **14**(1), pp. 16-24. 2012.

[11] M. Ghassemi, et al., "Unfolding physiological state: Mortality modelling in intensive care units," SIGKDD 2014: 75-84.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality," NeurIPS 2013: 3111-3119.

[13] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," arXiv:1301.3781. 2013.

[14] R. Collobert, J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," ICML 2008: 160-167.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," EMNLP 2014: 1532-1543.

[16] J. A. Minarro-Gimnez, O. Marin-Alonso, M. Samwald, "Exploring the application of deep learning techniques on medical text corpora," Studies in health technology and informatics, 205, 584-588, 2014.

[17] L. D. Vine, et al., "Medical semantic similarity with a neural language model," CIKM 2014: 1819-1822.

[18] W. Ling, et al., "Not all contexts are created equal: Better word representations with variable attention," EMNLP 2015: 1367-1372.

[19] T. Tran, T. D. Nguyen, D. Phung, S. Venkatesh, "Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)," Journal of biomedical informatics, 54, 96-105, 2015.

[20] Y. Choi, C. Y. I. Chiu, D. Sontag, "Learning low-dimensional representations of medical concepts," In Proc. AMIA Summits on Translational Science Proceedings, 41, 2016.

[21] O. Melamud, D. McClosky, S. Patwardhan, Bansal, "The Role of Context Types and Dimensionality in Learning Word Embeddings," HLT-NAACL, 1030-1040, 2016.

[22] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: graph-based attention model for healthcare representation learning," SIGKDD 2017: 787-795.

[23] L. Liu, F. Ruiz, S. Athey, and D. Blei, "Context selection for embedding models," NeurIPS 2017: 4816-4825.

[24] A.E. Johnson, et al., "MIMIC-III, a freely accessible critical care database." Scientific data, **3**, 160035, 2016.

[25] Z. Qiao, S. Zhao, C. Xiao, X. Li, Y. Qin, and F. Wang, "Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction," IJCAI 2018: 3520-3526.

[26] S. Wang, X. Li, L. Yao, Q. Z. Sheng, and G. Long, "Learning multiple diagnosis codes for ICU patients with local disease correlation mining," ACM Transactions on Knowledge Discovery from Data, 11(3), 31, 2017.

[27] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014.

[28] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv:1609.04747, 2016.

[29] Z. Yan, J. Fan, and J. Wang, "A collective neurodynamic approach to constrained global optimization," IEEE transactions on neural networks and learning systems, **28**(5), 1206-1215, 2017.

[30] G. Li, Z. Yan, and J. Wang, "A one-layer recurrent neural network for constrained nonconvex optimization," Neural Networks, **61**, 10-21, 2015.

[31] E. Choi, C. Xiao, W. Stewart, and J. Sun, "MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare," NeurIPS 2018: 4552-4562.