Attentive Dual Embedding for Understanding Medical Concepts in Electronic Health Records

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*—Electronic health records contain a wealth of information on a patients healthcare over many visits, such as diagnoses, treatments, drugs administered, and so on. The untapped potential of these data in healthcare analytics is vast. However, given that much of medical information is a cause and effect science, new embedding methods are required to ensure the learning representations reflect the comprehensive interplays between medical concepts and their relationships over time. Unlike one-hot encoding, a distributed representation should preserve these complex interactions as high-quality inputs for machine learning-based healthcare analytics tasks. Therefore, we propose a novel attentive dual embedding method called MC2Vec. MC2Vec captures the proximity relationships between medical concepts through a two-step optimization framework that recursively refines the embedding for superior output. The framework comprises a Skip-gram model to generate the initial embedding and an attentive CBOW model to fine-tune the embedding with temporal information gleaned from sequences of patient visits. Experiments with two public datasets demonstrate that MC2Vecs produces embeddings of higher quality than five state-of-the-art methods.**

***Index Terms*—medical concept embedding, attention mechanism, med2Vec, dual embedding**

I. Introduction

Today, most healthcare information systems store their data as electronic healthcare records (EHRs). Each EHR is a sequential record of a patients healthcare visits, where each visit is logged as a set of medical entities and concepts [2]. To ensure there is no ambiguity about what the data means, the medical concepts are recorded as codes following a set of standardized coding systems. There are codes, and coding systems, for diagnoses, medical procedures, drugs administered, and so on. Developed by medical experts, these coding systems are based on straightforward tree hierarchies that reflect the basic taxonomy of our current medical knowledge. Tree structures are easy for humans to understand and maintain, while they are good at representing simplistic vertical relationships, they are not good at representing the intricate complexities of horizontal and time-aware relationships. In EHRs, many medical concepts co-occur creating a much richer picture than a simple tree-based hierarchy can possibly illustrate. As such, there is a wealth of knowledge locked within these codes that, when revealed by healthcare analytics, can be put to purpose for example, for diagnoses prediction [1], [22], [25], predicting inpatient mortality [9], or the expected length of a hospital stay after admission [9].

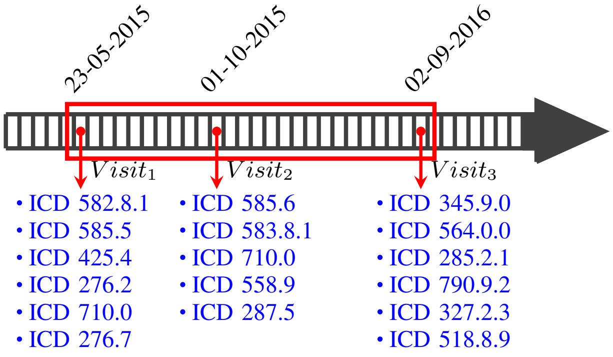

Fig. 1. An example segment of one patient's healthcare journey

The multi-level structure of an EHR comprises three layers: the patient, the visits associated with that patient, and the medical concepts associated with that visit. A patients healthcare journey, namely a patient journey, is a sequence of visits each occurring at a different time stamp. Fig. 1 shows an example segment of one patients journey from an EHR [5], [25]. The medical concept codes here are codes from the International Classification of Diseases (ICD).

Natural language processing (NLP) uses a similar multi-level structure with its document, sentence, and word structures, i.e., a document comprises a sequence of sentences, and each sentence is a bag of words. These parallels mean healthcare analytics may be able to borrow some of NLPs useful tools. However, there are some differences between the two domains.

- The visits in a patient journey are sequential but intermittent, while the sentences in a document are simply sequential.
- In the bag of medical concepts, one code is dominant, called the principle. All the other codes are considered to be of equal importance with no sequential relationships between them. Whereas, the words in a sentence have a sequential relationship, and all are treated as being equally important.
- Each medical code in a bag is unique, while a sentence may include repeated words.

Hence, drawing out the semantic information in EHRs for analytics and applications requires an embedding method. One-hot encoding a medical concept will generate a sparse high-dimensional vector, so a more straightforward solution is to use word embedding to create representations of the medical concepts for learning [7], [12], [14], [15]. These approaches have been shown to improve performance in a range of healthcare applications [16], [17], [19], [20], [22]. Extending this idea, Choi et al. and Edward et al. [1], [31] proposed multi-level representation learning to embed the visits and medical concepts simultaneously by using the sequential order of visits and the co-occurrence of medical concepts. Cai et al. [5] proposed a CBOW-based embedding method for medical concepts, enhanced with an attention mechanism that captures temporal information about the visits. The basic idea of their approach is to split the sequence of a patients visits into a number of discrete time units. Then, an attention mechanism captures both the sequence information and the time-aware information. These methods have merit. However, the time units are divided into fixed sizes, which is impractical because different diagnoses or treatments might have a different awareness of time. Moreover, the length of the time units is sensitive – a long unit might cause information loss by placing several visits into one time unit; a short unit may drastically increase dimensionality, causing an explosion in the attention mechanism. A better method is needed.

Our solution is to embed a medical concept into a vector with a novel attentive dual embedding technique, which fully leverages the information in multi-level EHRs. This dual embedding model, called MC2Vec (medical concept to vector), is controlled by a novel loss function designed to satisfy three objectives: 1) the target medical concept can accurately predict the context of the concept, i.e., one-to-N embedding; 2) the context surrounding the medical concept can accurately predict the target concept, i.e., N-to-one embedding; and 3) the attention mechanism can accurately attend the temporal sequence information. Then, the optimal solution, i.e., the embedding result, is discovered through a two-step optimization procedure. First, the medical concept is converted into a representation using one-hot encoding. Skip-gram is then used to embed the medical concept by taking that concept and using it to predict the surrounding context. These embedding results are generated as a one-to-N embedding. This one-to-N embedding forms a representation of the medical concept, which is used to train an attentive CBOW model that fine-tunes the embedding into an N-to-one embedding. These two steps are conducted recursively until the optimal embedding is produced.

A summary of this papers contributions to medical concept embedding include:

- a novel dual embedding method that fully leverages the information in EHRs with a new loss function that optimizes the workflow between two embedding models;
- an attentive CBOW method that captures temporal information in a flexible way and with less information loss due to its attention to time intervals; and
- a practice-driven method of embedding medical concepts that achieves start-of-the-art performance with two public datasets.

The rest of this paper is organized as follows. In Section II, we briefly discuss some related work. Then, the details of our method are presented in Section III. In Section IV, we demonstrate the results of experiments conducted on real-world public datasets. Lastly, we conclude our study in Section V along with our intentions for future work.

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 networks has only become a hot research topic since 2003 [3], [7], [12]–[15]. CBOW and the Skip-gram model [12], [13] are among two of the model families, that were introduced to compute continuous vector representations of words from very large data sets. Each is based on the assumption that the order of words or a words context do not influence the projection of the target word. However, recently, some scholars have begun to discover that sequence and context do matter. For example, Melamud et al. [21] explored the impact of context with the Skip-gram model, finding that weighting for context can improve performance with extrinsic tasks. Similarly, Liu et al. [23] show that conditioning a target word on a subset of contexts improves both the quality of the embedding and the predictions. Ling et al. [18] extended CBOW by incorporating an attention model that considers contextual words and their positions relative to the predicted word, which results in better representations. Each of these advancements has proven effective in the field of NLP but, as discussed in Section I, the differences between documents and patient journeys means these embedding models cannot be directly applied to medical concepts in EHRs without information loss or reduced performance.

B. Medical Concept Embedding

Borrowing ideas from word representation models [12], [13], researchers in the healthcare domain have recently explored the possibility of creating representations of medical concepts. Much of this research has focused on the Skip-gram model. For example, Minarro-Gimnez et al. [16] directly applied Skip-gram to learn representations of medical text, and Vine et al. [17] did the same for UMLS medical concepts. Choi et al. [20] went a step further and used the Skip-gram model to learn medical concept embeddings from different data sources, including medical journals, medical claims, and clinical narratives. In other work [1], Choi et al. developed the Med2Vec model based on Skip-gram to learn concept-level and visit-level representations simultaneously. The shortcoming of all these models is that they view EHRs as documents in the NLP sense, which means temporal information is ignored.

Attention mechanisms are a more recent introduction to healthcare analytics [8]. Choi et al. [22] proposed a graph-based attention model that learns representations of medical concepts from medical ontologies. Rajkomar et al. [9] applied an attention-based time-aware neural network model to predict patient outcomes, and Cai et al. [5] proposed MCE (Medical Concept Embedding) as a way to integrate time information into an attention model to embed medical concepts. Our work departs from Cai et al. [5] in that MC2Vec attends the time intervals between visits, and the context window is not based on time units but rather on temporal windows.

III. The Proposed Model

This section starts by introducing some definitions of medical concepts and the related notations. Then, we briefly introduce the basic units of medical concept embedding. The final subsection describes the proposed attentive dual embedding method.

A. Preliminaries

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

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

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

Definition 4 (Temporal Interval): Temporal interval refers to time difference between two visits in a patient journey, defined 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 concepts *C* to a real-valued dense vector of dimension *d*.

B. Basic Units of Medical Concept Embedding

The most straightforward embedding method is to adapt a classic embedding model [12], [13], such as CBOW or Skip-gram, to tackle medical concept embedding tasks. The basic idea is to generate training samples from EHRs by selecting one medical concept as the target vector and its co-occurring or correlated medical concepts as the context. In the medical concept version of CBOW, the representations are learned by constructing a neural network classification model that uses a context vector comprising multiple medical concepts to predict the target word. This is also known as N-to-one embedding. In Skip-gram, rather than predicting the target word based on the context, each target vector is used as an input to predict a context vector. This is known as one-to-N embedding.

Given EHRs's multi-level structure, we have elected to extract training samples using medical concepts that co-occur in a sequence of visits. Each visit V_t is a bag of medical concepts $\{c_1, c_2, ...\}$. Each c_i in the bag is transformed into a target vector, and its context vector $H = \{c_k, c_l, c_m, c_n, ...\}$ is constructed by randomly sampling medical concepts from this bag. Sometimes, the first concept in the bag will be allocated a higher probability of being sampled because, traditionally, the first concept is the principle code, i.e., the disease, procedure, drug, etc. that dominated the visit. At other times, a sliding window might be used to sample related medical concepts based on the assumption that medical doctors commonly annotate highly correlated concepts together. Which alternative is chosen is based on an empirical analysis of the dataset. For simplicity, we have chosen to outline the sliding window method in our description of the model.

1) CBOW-based medical concept embedding: The objective of CBOW is to maximize the average log probability of the occurrence of a target vector c given a context vector H. For a given visit, the objective function can be defined as a 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 comprising the medical concepts sampled from the sliding window.

2) Skip-gram-based medical concept embedding: The objective of the Skip-gram model is to maximize the average log probability of predicting a context vector using a target vector. This objective function is defined as

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

where K is the total number of medical concepts in the context vector, and c_j' is a medical concept in the 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 a generalized softmax function, regardless of whether the CBOW-based N-to-one embedding or the Skip-gram-based one-to-N embedding is used. The general definition is: use an input vector c_I to predict c_O. The probability of prediction is:

$$p(c_O|c_I) = \frac{\exp\{{v_{c_O}^{'}}^{T} v_{c_I}\}}{\sum_{c=1}^{|C|} \exp\{{v_c^{'}}^{T} v_{c_I}\}}, \tag{3}$$

where v_c and $v_c^{'}$ are the "input" and "output" vector representations of c, and $|C|$ is size of the medical concept vocabulary. For CBOW, $v_{c_I} = (1/2k) * \sum_{c_j \in H_t} v_{c_j}$.

3) Negative Sampling: The formulation of 3 is impractical for computation because the cost of computing $\bigtriangledown \log p(c_O|c_I)$ is proportional to $|C|$, which is often large. To reduce the computational complexity, the Word2Vec model uses negative sampling to replace every $\log p(c_O|c_I)$ term in CBOW and Skip-gram objectives. Rather, the objective is to maximize

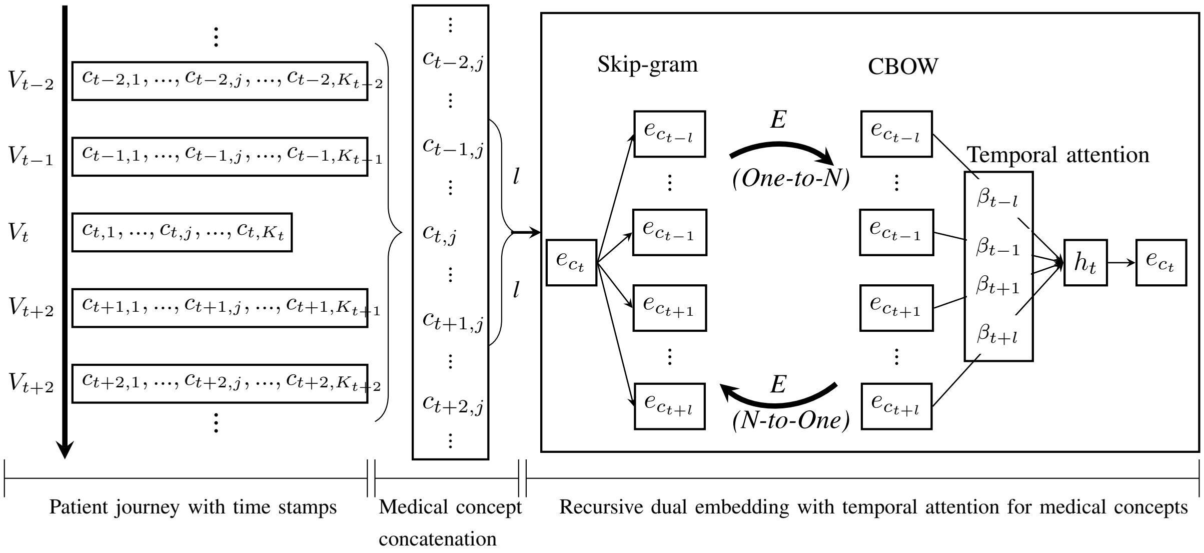
V_{t-2}
$c_{t-2,1}, ..., c_{t-2,j}, ..., c_{t-2,K_{t+2}}$
V_{t-1}
$c_{t-1,1}, ..., c_{t-1,j}, ..., c_{t-1,K_{t+1}}$
V_t
$c_{t,1}, ..., c_{t,j}, ..., c_{t,K_t}$
V_{t+2}
$c_{t+1,1}, ..., c_{t+1,j}, ..., c_{t+1,K_{t+1}}$
V_{t+2}
$c_{t+2,1}, ..., c_{t+2,j}, ..., c_{t+2,K_{t+2}}$
$c_{t-2,j}$
$c_{t-1,j}$
$c_{t,j}$
$c_{t+1,j}$
$c_{t+2,j}$
l
l
Skip-gram
CBOW
e_{c_t}
$e_{c_{t-l}}$
$e_{c_{t-1}}$
$e_{c_{t+1}}$
$e_{c_{t+l}}$
E
(One-to-N)
E
(N-to-One)
Temporal attention
β_{t-l}
β_{t-1}
β_{t+1}
β_{t+l}
h_t
e_{c_t}
Patient journey with time stamps
Medical concept concatenation
Recursive dual embedding with temporal attention for medical concepts

Fig. 2. The dual embedding model for medical concepts with temporal attention. There are three stages. Stage 1 decomposes a patient journey into sequential visits. Stage 2 concatenates the concepts in each visit into a patient vector. Stage 3 is the dual embedding step for the concepts in a temporal window *l*, which integrates the three components of Skip-gram, CBOW, and temporal attention (where *E* represents the embedding parameters).

$$J = \log \sigma({v'_{c_O}}^T v_{c_I}) + \sum_{i=1}^{r} \mathbb{E}_{c_i \sim P(c)}[\log \sigma(-{v'_{c_i}}^T v_{c_I})], \quad (4)$$

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

C. Attentive Dual Embedding Approach

One of the unique elements that separate standard documents from patient journeys is that medical concepts have temporal relationships, whereas words do not. This temporal information is important for the embeddings. Therefore, we propose an attentive dual embedding method that comprises one-to-N and N-to-one embeddings so as to capture multiple views of the comprehensive semantic relationships in an EHR as well as the temporal information.

1) Architecture: MC2Vec has three parts: a) a patient journey with time stamps; b) the medical concept concatenation; and c) the attentive dual embedding. The architecture of the framework is illustrated in Fig. 2.

a) Patient journey with time stamps: The patient journey is split into *M* visits, i.e. $V_t = \{c_{t,1}, ..., c_{t,j}, ..., c_{t,K_t}\}$, where *t* is the tine if patient visit, and each medical concept *c* is associated with a time stamp *t*.

b) Medical concept concatenation: To generate a context and target concept for MC2Vec, the visits in the patient journey are concatenated according to their temporal sequence into a vector of medical concepts with time stamp. For example, suppose 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}, a temporal window size of *l*, and target concept c_t, we first leverage Skip-gram with a temporal window to learn the embedding parameters *E* of the medical concepts over the context as *one-to-N* embedding. Then we use a *one-to-N* embedding of *E* and temporal attention to learn the medical concept representations in the same window with an attentive CBOW model. The embedding produced is *N-to-one*. *One-to-N* works like an expectation step in the (EM) algorithm [6] as it fixes the embedding parameters of the target concept c_t to optimize the embedding parameters of its contextual concepts. Similarly, *N-to-one* is like the maximization step of the EM algorithm in that the embedding parameters of the target concept c_t are optimized by fixing the embedding parameters of the context concepts. By sliding a temporal window *l* over J_{vec} to view a different target concept c_t, *one-to-N* and *N-to-one* mutually reinforce each other to learn optimized embeddings.

The dual embeddings consist of three components: Skip-gram, CBOW, and temporal attention. On the 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, MC2Vec can improve the quality of medical concept embeddings by capturing temporal distributions.

d) Unified training: A single unified framework for generating an optimized representation of a medical concept can be built by summing the objective functions of *one-to-N* (Skip-gram) and *N-to-one* (attentive CBOW), i.e.,

$$\max_{E} J_{One2N} + J_{N2One} \quad (5)$$

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

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

where E denotes the embedding parameters, c_t is the target medical concept, h_t is the weighted context of c_t, c_x is 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, the medical concept embeddings can be learned from the same temporal window.

2) Temporal Attention: To capture the semantic relationships between medical concepts over time, we developed a temporal attention mechanism that is able to learn non-uniform attention weights in a temporal window. Specifically, the embedding results from the Skip-gram model form the inputs to the attentive CBOW embedding model, and the context vector is calculated by non-uniformly 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 window, $\log(2l+1)$ and $\log(\sum_{e_{c_i} \in H_t} \beta_i^2)$ are scalars to the weighted sum $\sum_{e_{c_i} \in H_i} \beta_i e_{c_i}$.

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

To calculate 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}. A k dimensional projection of the embedding is defined by learning a $k \times d$ dimensional matrix P and multiplying it to get the k scalars $p_{1,j}, ..., p_{k,j}$ for $e_{c_j} \in H_t$. The attribution logits are defined as

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

Thus, the model learns to pay more attention to the temporal intervals, and the medical concept representations are improved by identifying the time intervals between related visits and, in turn, capturing more accurate related target-context pairs.

3) Model Parameters and Complexity: Model training is conducted through Adam [27], one of the gradient descent optimizers [27]–[30], with the default recommended parameters. The only additional computation required beyond Skip-gram and CBOW is the temporal attention. Each operation for computing the attention weight multiplies P with e_{c_j}. Hence, the added computational complexity is related to the temporal attention window k, which is discussed in further detail in Section IV. The details of MC2Vec are shown in Algorithm 1 below.

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

We evaluated the quality of MC2Vecs embedding results with two public datasets on a machine learning clustering task.

A. Dataset Descriptions

Details of the two datasets used follow.

a) CMS: is a publicly available[1] synthetic claims dataset, which includes four types of files: inpatient, outpatient, carrier, and beneficiary summary. We chose to use a subset of the inpatient files for the period 2008 to 2010.

b) MIMIC III: [24] is an open-source, large-scale, de-identified dataset of EHR records for ICU patients. The dataset mainly consists of clinical logs for patients admitted to critical care units with serious conditions. The diagnosis codes are derived from the International Classification of Diseases (ICD9) system[2].

The statistical information for both datasets is listed in Table 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 ground truths for the clustering task were selected from two well-organized ontologies: the ICD9 standards and the Clinical Classifications Software (CCS) [3]. ICD9 has a

[1] https://www.cms.gov

[2] http://www.icd9data.com

[3] https://www.hcup-us.ahrq.gov

hierarchical structure [26], as shown in Fig. 3. For example, the first three numbers of all codes ranging from 460 to 519 are classified as diseases of the respiratory system, which is one of 19 categories. We used the high-level nodes as the clustering labels. Both the MIMIC III and the CMS datasets contained all 19 categories of disease. These ground truths are denoted as **ICD**. CCS provides a way to classify ICD9 diagnosis codes and other medical procedures into 285 broad but mutually exclusive diagnoses and procedure groups for statistical analysis and reporting [4]. Examples of the CCS diagnosis categories are shown in Table II. The MIMIC III dataset contained 265 of these categories and CMS contained 274. These ground truths are denoted as **CCS**.

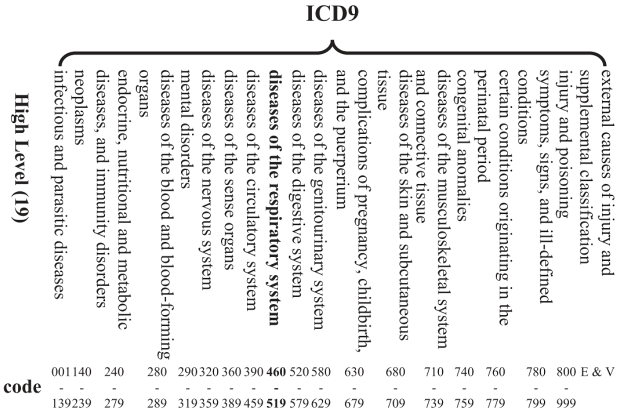
ICD9
High Level (19)
infectious and parasitic diseases
neoplasms
endocrine, nutritional and metabolic diseases, and immunity disorders
diseases of the blood and blood-forming organs
mental disorders
diseases of the nervous system
diseases of the sense organs
diseases of the circulatory system
diseases of the respiratory system
diseases of the digestive system
diseases of the genitourinary system
complications of pregnancy, childbirth, and the puerperium
diseases of the skin and subcutaneous tissue
diseases of the musculoskeletal system and connective tissue
congenital anomalies
certain conditions originating in the perinatal period
symptoms, signs, and ill-defined conditions
injury and poisoning
external causes of injury and supplemental classification
code
001 - 139
140 - 239
240 - 279
280 - 289
290 - 319
320 - 359
360 - 389
390 - 459
460 - 519
520 - 579
580 - 629
630 - 679
680 - 709
710 - 739
740 - 759
760 - 779
780 - 799
800 - 999
E & V

Fig. 3. The hierachical structure of ICD9.

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

A brief description of the five state-of-the-art embedding methods chosen as baseline comparisons is provided below.

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

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

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

d) ***med2vec*** *[1]:* is a multi-level embedding model that creates d) embedding medical concepts and visits simultaneously.

e) ***MCE*** *[5]:* is a CBOW model with time-aware attention that embeds medical concepts with temporal information.

We also tested the subcomponents of MC2Vec independently for further comparison, as follows:

f) ***CBOW_Attn****:* is based on CBOW but integrates the temporal intervals of sequential visits into an attention model to learn representations of medical concepts. This is the attentive CBOW component of MC2Vec.

g) ***Sg_CBOW****:* is a vanilla version of our dual embedding model MC2Vec without the attention mechanism.

h) ***MC2Vec****:* is our proposed attentive dual embedding model for medical concepts that integrates Skip-gram and attentive CBOW to learn representations of medical concepts.

All datasets were preprocessed to remove infrequent medical concepts with an empirically set threshold of 5. Following the original Word2Vec [12], [13], the same negative sampling strategy was used for Skip-gram and CBOW, CBOW_Attn, Sg_CBOW and MC2Vec. The number of negative samples for both MIMIC III and CMS was set to 10 and 5, respectively. All models were trained for 10 epochs with MIMIC and for 5 epochs with CMS. The dimension d of the medical concept embeddings was set to 100. The temporal window for MC2VEC was empirically set to 9 for both datasets.

D. Results

This section presents the results of the clustering task with all models. We used K-Means as the clustering algorithm and evaluated the learned representations against the two sets of ground truths in terms of normalized mutual information (NMI). The results appear in Table III. The best results appear in bold.

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: As the results show, MC2Vec delivered the best results against both ground truths with both datasets at a window of 9. We attribute this performance

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

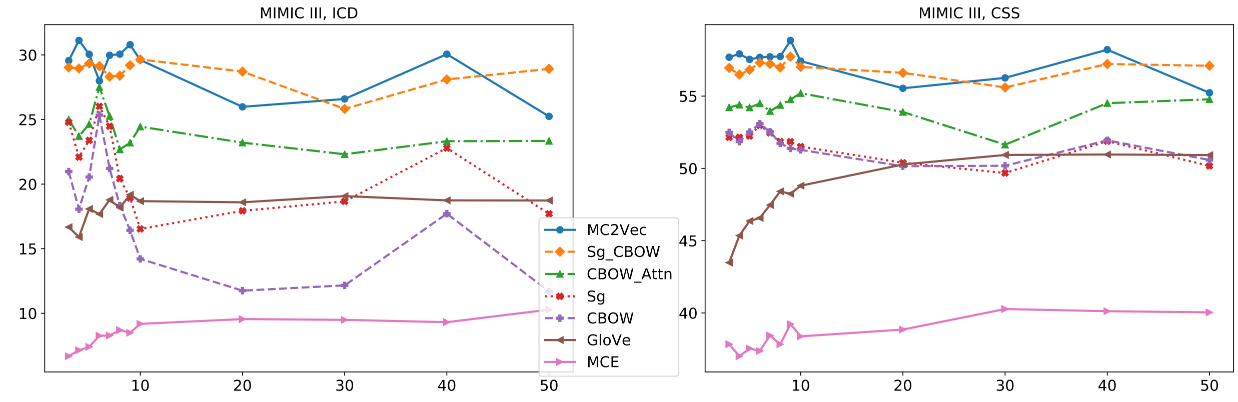

Fig. 4. NMI (%) of the models on MIMIC III w.r.t. ground truth ICD and CCS. The window size varies from 3 to 50.

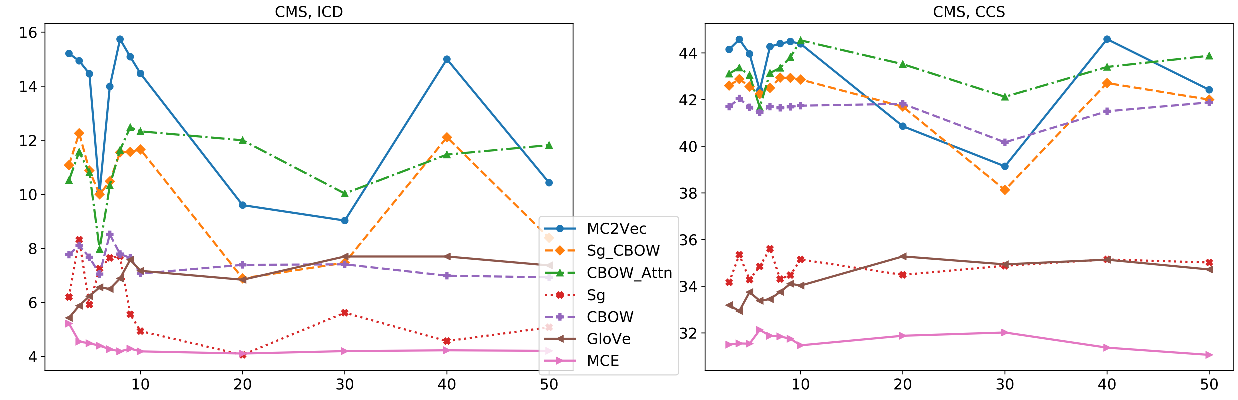

Fig. 5. NMI (%) of the models on CMS w.r.t. ground truth ICD and CCS. The window size varies from 3 to 50.

to the introduction of the dual embedding model, which incorporates the temporal attention into the model. In turn, the model learns better embeddings of the medical concepts. Notably, the performance of the two components of MC2VEC, Sg_CBOW, and CBOW_Attn, was also very competitive. All models performed better when compared to the CCS ground truth than ICD. CCS has a well-organized ontology based on expert knowledge, which may explain this result.

b) Performance of varying window sizes: To evaluate the effects of the context window, we varied the size of the window from 3 to 50 and compared the full version of MC2Vec with the five of the six baselines. Med2Vec was omitted because this model does not include a parameter for window size. The results with MIMIC III on the same clustering task as above are summarized in Fig. 4, and Fig. 5 shows the results with the CMS dataset.

For most models, performance decreased as the window size increased due to the additional noise a larger window size introduces. However, GloVe makes use of global co-occurrences, and MCE has greater temporal scope, so neither of these models were sensitive to window size. In fact, these two models showed better performance as the window size grew. MC2Vec and Sg_CBOW showed competitive performance and consistently produced better results than the rest models in terms of NMI. This demonstrates that integrating the two embedding models does capture the relationships between medical concepts in a more comprehensive way.

Turning to the results with the CMS dataset in Fig. 5, MC2Vec outperformed the other baselines with a skip window size of not more than 10, after which CBOW_Attn took over. This interesting result indicates that attention brings benefit to the quality of the embeddings.

The GloVe, Skip-gram, and MCE models remained relatively stable no matter the window size. The other models reached the local minimum at a skip window of 6 with their best performance at 8.

c) Influence of the Attention Window k: Fig. 6 shows the change in MC2Vecs performance with different attention window sizes. Here, k was varied from 10 to 500. The two vertical axes represent the range of results.

The results show the best performance with MIMIC III at an attention window size of 300 for both ICD and CCS, but performance dropped quickly once the window size reached 300. This is due to data sparsity in the EHRs in the period 2001 to 2012. With the CMS dataset, MC2Vecs best performance occurred at an attention window of 100. This demonstrates MC2Vecs effectiveness at classifying dense, large-scale datasets.

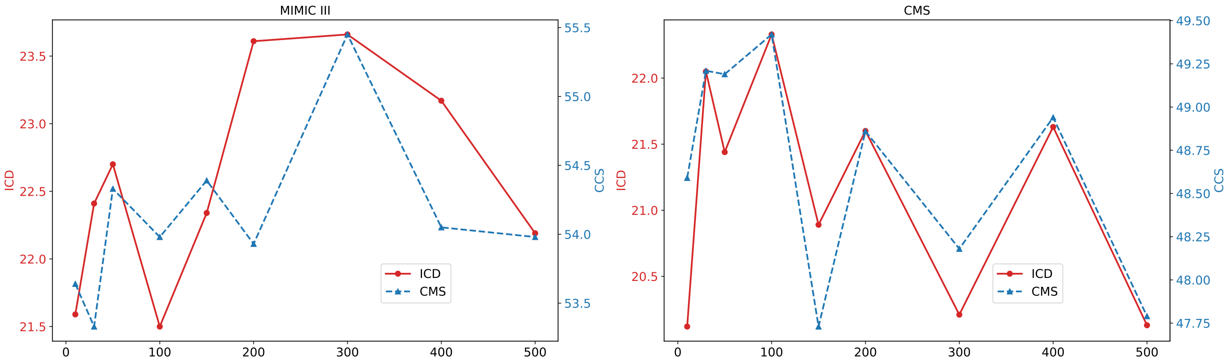

Fig. 6. NMI (%) on MIMIC III and CMS by varying attention windown *k* from 10 to 500.

V. Conclusion

This paper presents an attentive dual embedding method, called MC2Vec, that captures multiple views of the comprehensive relationships between medical concepts. The model comprises Skip-gram, which generates the initial embedding and attentive CBOW, which refines the embeddings with temporal information. The two models operate in a recursive manner to produce superior embeddings for use with machine-learning-based healthcare analytics. Comparative experiments with five state-of-the-art baselines on two public datasets show MC2Vec produces significantly better quality representations for clustering tasks. In next step, we plan to build patient journey graph [32]–[34] to learn medical concept embedding.

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 J Biomed Health Inform, **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,” Stud Health Technol Inform, 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),” J Biomed Inform, 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,” TKDD, 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 Trans Neural Netw Learn Syst, **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.

[32] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, ”A comprehensive Survey on Graph Neural Networks,” arXiv:1901.00596, 2019.

[33] S. Pan, J. Wu, X. Zhu, C.,Zhang, and Y. Wang, ”Tri-party Deep Network Representation,” Network, **11**(9), 12, 2016.

[34] X. Shen, S. Pan, W. Liu, Y.S. Ong, Q.S. Sun, ”Discrete Network Embedding,” IJCAI 2018: 3549-3555.