Medical Concept Representation Learning from Multi-source Data

Tian Bai[1] , **Brian L. Egleston**[2] , **Richard Bleicher**[2] and **Slobodan Vucetic**[1]
[1]Department of Computer and Information Sciences, Temple University, USA
[2]Fox Chase Cancer Center, USA
tue98264@temple.edu, {brian.egleston, richard.bleicher}@fccc.edu, vucetic@temple.edu

Abstract

Representing words as low dimensional vectors is very useful in many natural language processing tasks. This idea has been extended to medical domain where medical codes listed in medical claims are represented as vectors to facilitate exploratory analysis and predictive modeling. However, depending on a type of a medical provider, medical claims can use medical codes from different ontologies or from a combination of ontologies, which complicates learning of the representations. To be able to properly utilize such multi-source medical claim data, we propose an approach that represents medical codes from different ontologies in the same vector space. We first modify the Pointwise Mutual Information (PMI) measure of similarity between the codes. We then develop a new negative sampling method for word2vec model that implicitly factorizes the modified PMI matrix. The new approach was evaluated on the code cross-reference problem, which aims at identifying similar codes across different ontologies. In our experiments, we evaluated cross-referencing between ICD-9 and CPT medical code ontologies. Our results indicate that vector representations of codes learned by the proposed approach provide superior cross-referencing when compared to several existing approaches.

1 Introduction

Medical claims are files created by medical providers for billing purposes to summarize the services provided to patients. A medical claim includes a list of medical codes, which describe patient diagnosis and treatment. For example, the International Classification of Diseases (ICD) and Current Procedural Terminologies (CPT) medical code ontologies contain tens of thousands of codes for diseases and medical procedures. While ICD ontology contains alphanumeric codes for both diagnoses and procedures associated with patient treatment, CPT ontology is used to solely describe treatment. Beyond their primary purpose in billing, medical claims have been widely used in healthcare research for exploratory analysis and predictive analytics [Bai and Vucetic, 2019; Bai *et al.*, 2018b].

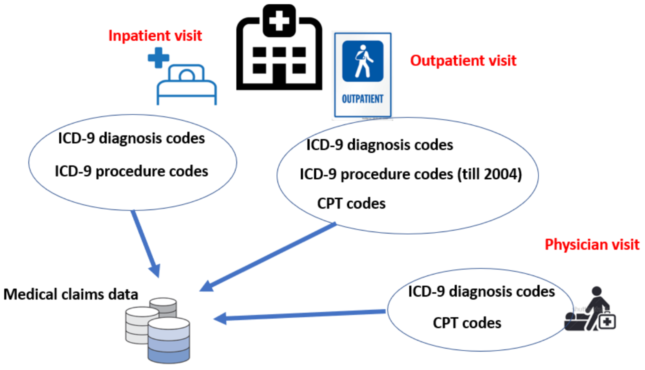

Figure 1: Different providers use different ontologies to record patient diagnosis and treatment. Those multi-source medical claims are stored in a medical claim database.

To improve analysis of medical claim data, recent medical informatics research has focused on finding vector representations of medical codes, in which each code is represented as a vector and where related codes are neighbors in the vector space. Based on the distributional hypothesis, the neighboring codes in the vector space are those that occur in the similar contexts. As a representative of this line of work, [Choi *et al.*, 2016b] used word2vec algorithm [Mikolov *et al.*, 2013] to learn the vector representations of medical codes using longitudinal medical records data and show that the related codes indeed obtain similar vector representations. Other examples also illustrated the benefits of vector representations: [Choi *et al.*, 2016a] used a multi-layer perceptron to learn representations for predicting future medical codes and clinical risk groups. [Choi *et al.*, 2017] incorporated ontological knowledge of medical codes into an attention model in order to learn improved code representations. [Bai *et al.*, 2017; Bai *et al.*, 2018a] modified word2vec algorithm in order to learn joint embeddings of clinical words and medical codes. [Cai *et al.*, 2018] incorporated a time-aware attention mechanism into CBOW word2vec model, which takes progression patterns of different diseases into consideration.

However, the direct application of word2vec and related algorithms might not be appropriate when dealing with medical claims obtained from multiple types of providers, in which every provider may decide to use different coding ontologies. To understand the underlying issue, let us consider medical claims from Medicare, the U.S. federal insurance system for senior citizens, which is a popular source of data for healthcare research. As shown in Figure 1, the format of medical claims varies with the type of service [Warren *et al.*, 2002]. In particular, the 3 main types of medical claims used by Medicare recognize 3 types of patient interactions with the medical system: (1) Inpatient, summarizing services requiring a patient to be admitted to a hospital, (2) Outpatient, summarizing services which do not require a hospital stay, and (3) Carrier, which refer to services by non-institutional providers such as physicians or registered nurses. Each inpatient claim is one summarized record per hospital admission and includes up to 25 ICD diagnosis codes and 25 ICD procedure codes. Carrier claims, on the other hand, use CPT codes to record procedures and ICD diagnosis codes to justify the reason for the service. Until 2004, outpatient claims contained both ICD diagnosis and procedure codes, while after 2004 the ICD procedure codes were dropped. This information is summarized in Figure 1.

Let us explain why the direct use of word2vec on a set of medical claims coming from different types of providers, such as in Medicare, is problematic. As shown in [Levy and Goldberg, 2014], word2vec's Skip-gram with negative sampling algorithm is implicitly performing a factorization of a Pointwise Mutual Information (PMI) matrix [Turney and Pantel, 2010]. The (i, j)-th element of the PMI matrix reflects how much presence of code j in a claim increases the probability of seeing code i in the claim. If the codes i and j are from different ontologies, a straightforward calculation of the probabilities could lead to misleading results. In particular, the PMI values for pairs of codes from different ontologies could be severely underestimated, which could result in suboptimal vector representation of codes.

In this work we propose a modification that correctly estimates the PMI scores between medical codes from different ontologies. Following this modification, we also propose an improved version of negative sampling method inside the word2vec algorithm. It is worth noting that recent papers [Wang *et al.*, 2018; Grbovic and Cheng, 2018; Cai and Wang, 2018] also observed that traditional negative sampling has drawbacks in some applications. For example, [Wang *et al.*, 2018] incorporated Generative Adversarial Networks (GAN) into negative sampling mechanism in order to generate high-quality negative samples. [Grbovic and Cheng, 2018] replaced negative sampling probability with a set of probabilities suited for different subsets of embedding objects. However, previous papers lack theoretical justification for the proposed approaches. One of the main contributions of our paper is in explaining that deficiency of negative sampling on multi-source data could be traced to the PMI formula. We propose how to modify the PMI formula and develop the corresponding negative sampling mechanism. To demonstrate the effectiveness of the proposed algorithm, we jointly learn vector representations of ICD and CPT codes using a large Medicare medical claim dataset related to breast cancer. Since ICD procedure codes and CPT codes are both encoding medical procedures, we were able to observe how close in the vector space are the codes from the two ontologies that represent highly similar procedures. We note that this particular evaluation is closely related to the long-standing problem of cross-referencing [Brouch, 2004; Topaz and Shafran-Topaz, 2013; Schulz *et al.*, 1998; Butler, 2007], that deals with creation of mappings between different coding ontologies, such as ICD-9 and CPT, ICD-9 and ICD-10, or ICD-9 and SNOMED.

In the next section we provide background information about PMI metric and word2vec algorithm, and describe the proposed modifications. Then we present the experimental results.

2 Method

2.1 Problem Setup

Let us assume we are given a dataset of patient visits $S = \{s_1, s_2, ..., s_{|S|}\}$, where $|S|$ is the number of visits. Each visit s_t consists of a set of codes summarizing the visit. Let us denote the set of all codes $C = \{c_1, c_2, ..., c_{|C|}\}$, where $|C|$ is the number of codes. The objective is to find vector representation of codes, such that each code c_i is represented as a K-dimensional vector V_i. A good vector representation would place related codes in the vicinity in the K-dimensional vector space. In the following subsection we will overview two popular methods from the NLP community that have already been used with success in medical informatics.

2.2 PMI and Skip-gram

Pointwise mutual information (PMI) [Turney and Pantel, 2010] measures how much co-occurrence of two codes in claims deviates from the predicted co-occurrence if they were independent. The PMI between codes c_i and c_j is defined as

$$PMI_{ij} = PMI(c_i, c_j) = \log \frac{P(c_i|c_j)}{P(c_i)}, \qquad (1)$$

where $P(c_i|c_j)$ is the conditional probability of seeing code c_i in a claim if c_j is already in the claim and $P(c_i)$ is the marginal probability of seeing c_i co-occurring with any code. The standard way to calculate those probabilities is to count code co-occurrence in the following way. Let us denote the count of times code c_i and c_j co-occur in the claims as n_{ij}. Then, denote with n_i the number of times code c_i co-occurs with any code: $n_i = \sum_{c_j \in C} n_{ij}$. Finally, denote with n the number of times two codes co-occur: $n = \sum_{c_i \in C} n_i$. Given those values, we can estimate the probabilities as $P(c_i|c_j) = \frac{n_{ij}}{n_j}$ and $P(c_i) = \frac{n_i}{n}$ and the PMI becomes

$$PMI_{ij} = \log \frac{n_{ij} \cdot n}{n_i \cdot n_j}. \qquad (2)$$

Using PMI scores, we can create a positive PMI (PPMI) matrix M of dimension $|C| \times |C|$, whose element $M_{ij} = \max(PMI_{ij}, 0)$. [Levy and Goldberg, 2014] propose to measure similarity between codes c_i and c_j by calculating the cosine similarity between i-th and j-th row of M. Since the PPMI matrix is often sparse, noisy, and large, it is often a good idea to apply Singular Value Decomposition (SVD) to factorize this matrix as $M = U\Sigma V$ and use the first K left eigenvectors (truncated version of U which retains the first K columns) for code representation. In particular, the i-th row of the truncated matrix U becomes the low-dimensional vector representation of code c_i.

Instead of using SVD on the PPMI matrix, which can be very costly when the code vocabulary is large, the recently proposed Skip-gram method is a space-efficient alternative to learn low dimensional representations of medical codes [Choi *et al.*, 2016b]. Skip-gram relies on scanning the visits sequentially. For every code c in t-th visit s_t, all other codes in the same visit $c' \in s_t$ are chosen as its context. The output of Skip-gram are two matrices: code matrix $V \in \mathbb{R}^{|C| \times K}$ and context code matrix $W \in \mathbb{R}^{|C| \times K}$. Code $c_i \in C$ is then associated with the i-th row of V and the i-th row of W. [Mikolov *et al.*, 2013] propose Skip-gram with negative sampling (SGSN)[1] to maximize the objective function

$$\begin{aligned} l = &\sum_{t=1}^{|S|} \sum_{c \in s_t} \sum_{c' \in s_t} (\log p(S=1|c,c') \\ &\qquad + k\mathbb{E}_{c_N \sim P}\left[\log(1 - p(S=1|c,c_N))\right]) \\ = &\sum_{c_i \in C} \sum_{c_j \in C} n_{ij} (\log \sigma(V_i \cdot W_j) \\ &\qquad + k\mathbb{E}_{c_N \sim P}\left[\log \sigma(-V_i \cdot W_N)\right]), \end{aligned} \tag{3}$$

in which $p(S = 1|c_i, c_j)$ denotes the probability that c_i and c_j are observed co-occurring in the dataset S and is defined as a sigmoid function $\sigma(V_i W_j)$,

$$p(S = 1|c_i, c_j) = \sigma(V_i \cdot W_j) = \frac{1}{1 + e^{-V_i \cdot W_j}}, \tag{4}$$

$P(c_N) = \frac{n_N}{n}$ is a frequency distribution over the vocabulary from which c_N is drawn, and k is the number of randomly generated negative codes for each scanned and context code pair (c, c'). The first term of l represents the probability that the scanned code c and its context code c' are observed co-occurring in the dataset S and the second term of l represents the probability that the scanned code c and randomly drawn "negative" code c_N are not observed co-occurring in the dataset S. Code matrices V and W are learned by stochastic gradient algorithm. Rows of V are used as vector representations of codes. Interestingly, [Levy and Goldberg, 2014] shows that SGSN is implicitly factorizing a shifted PMI matrix, making a connection between the PMI and Skip-gram approaches.

2.3 PMI Learned from Multi-source Data

As we discussed in the Introduction, different types of providers might use different types of codes in their claims. On such data, PMI defined in (1) is likely to produce misleading results. To see why, let us consider PMI between an ICD-9 procedure code and a CPT code. Even if the 2 codes are highly related, they can only co-occur in outpatient claims prior to 2004 (see Figure 1) and their n_{ij} count will be low compared to the n_i and n counts, which would lead to a negative PMI value. The opposite effect would occur with PMI score between two ICD-9 procedure codes. If both c_i and c_j are ICD-9 procedure codes, since the presence of an ICD-9 procedure code c_j in the claim implies the type of the claim is not carrier or outpatient after 2004, $P(c_i|c_j)$ is likely to be larger than $P(c_i)$. Then PMI would be large even if the 2 codes are unrelated.

To reduce the negative impact of varying code coverage among different types of claims, we propose a modified PMI definition as described next. Let us first denote the set of all code types (*i.e.*, ICD-9 diagnosis codes, ICD-9 procedure codes, CPT codes) as $T = \{t_1, ..., t_{|T|}\}$, where $|T|$ is the number of code types. We define function $f : C \mapsto T$ which maps every code $c_i \in C$ to its type $f(c_i) \in T$. We then calculate the number of times code c_i co-occurs with another code of type t_j as $n_i^{t_j} = \sum_{c_k \in C, f(c_k) = t_j} n_{ik}$, and the number of times a code of type t_i co-occurs with a code of type t_j as $n^{t_i t_j} = \sum_{c_k \in C, f(c_k) = t_i} n_k^{t_j}$. The modified PMI, called the $typePMI$, is then defined as

$$typePMI_{ij} = \log \frac{n_{ij} \cdot n^{f(c_i) f(c_j)}}{n_i^{f(c_j)} \cdot n_j^{f(c_i)}}. \tag{5}$$

To illustrate the benefit of typePMI, let us consider ICD-9 diagnosis code "83942" (dislocation of sacrum) and ICD-9 procedure code "8839" (X-ray). In practice, X-ray is used to diagnose dislocation and the 2 codes are related. In our dataset, the standard PMI between these two codes is -0.65, implying they are slightly mutually exclusive, while the typePMI is 4.34, correctly reflecting that the codes are strongly co-occuring.

The typePMI can be be used to construct the typePPMI matrix with elements $M_{ij} = \max(typePMI_{ij}, 0)$. We propose **typeSVD**, which applies SVD to typePPMI matrix to obtain a low-dimensional representation of codes.

2.4 Skip-gram Model from Multi-source Data

In this section we propose **typeSkip-gram** algorithm and prove that it is implicitly factorizing a shifted typePMI matrix. The key change is to replace the negative sampling probability distribution $P(c_N)$ with a new one which takes the code type into consideration:

$$P^{f(c_i), f(c_j)}(c_N) = \begin{cases} 0 & \text{if } f(c_N) \neq f(c_j) \\ \frac{n_N^{f(c_i)}}{n^{f(c_i) f(c_j)}} & \text{if } f(c_N) = f(c_j) \end{cases} \tag{6}$$

Unlike Skip-gram objective of (3), the negative sampling probability of (6) is formed by counting code pairs whose first code has the same type as the scanned code c_i and the second code has the same type as context code c_j.

Like Skip-gram, typeSkip-gram model is trained in an online fashion: it sequentially scans through codes in all visits $S = \{s_1, s_2, ..., s_{|S|}\}$. For each code $c \in s_t$, code $c' \in s_t$ in the same visit is chosen as its context code. For a specific code c and its context code c', typeSkip-gram uses stochastic

[1] The original objective function of Skip-gram is the sum of logarithmic softmax function which is computationally expensive to optimize. In this paper we focus on its practical implementation through negative sampling.

gradient descent algorithm to maximize

$$\begin{aligned}&\log p(S=1|c,c')\\&+k\mathbb{E}_{c_N\sim P^{f(c),f(c')}}\left[\log(1-p(S=1|c,c_N))\right].\end{aligned}\tag{7}$$

Given the new objective function (7), using the negative sampling probability distribution $P(c_N)$ defined in (6), we have the following theorem:

Theorem 1 *Given the sample distribution defined in (6), the dot product of V_i and W_j, $M'_{ij}=V_i\cdot W_j$, is the shifted typePMI between code c_i and c_j when objective function (7) is maximized:*

$$M'_{ij}=V_i\cdot W_j=typePMI_{ij}-\log k,\tag{8}$$

in which k is a constant.

Proof: Similar to (3), the objective function of typeSkip-gram is

$$\begin{aligned}l=&\sum_{c_i\in C}\sum_{c_j\in C}n_{ij}(\log\sigma(V_i\cdot W_j))\\&+\sum_{c_i\in C}\sum_{c_j\in C}n_{ij}(k\mathbb{E}_{c_N\sim P^{f(c_i),f(c_j)}}\left[\log\sigma(-V_i\cdot W_N)\right]).\end{aligned}\tag{9}$$

Since $n_i^{t_k}=\sum_{c_j\in C,f(c_j)=t_k}n_{ij}$, the second term of (9) can be written as

$$\sum_{c_i\in C}\sum_{t_k\in T}n_i^{t_k}(k\mathbb{E}_{c_N\sim P^{f(c_i),t_k}}\left[\log\sigma(-V_i\cdot W_N)\right]).\tag{10}$$

Since

$$\begin{aligned}&\mathbb{E}_{c_N\sim P^{f(c_i),t_k}}\left[\log\sigma(-V_i\cdot W_N)\right]\\&=\sum_{c_N\in C,f(c_N)=t_k}\frac{n_N^{f(c_i)}}{n^{f(c_i)f(c_N)}}\log\sigma(-V_i\cdot W_N),\end{aligned}$$

we can explicitly express the expectation term in (10) as

$$\begin{aligned}&\sum_{c_i\in C}\sum_{t_k\in T}n_i^{t_k}(k\sum_{\substack{c_N\in C\\f(c_N)=t_k}}\frac{n_N^{f(c_i)}}{n^{f(c_i)f(c_N)}}\log\sigma(-V_i\cdot W_N))\\&=\sum_{c_i\in C}\sum_{t_k\in T}\sum_{\substack{c_N\in C\\f(c_N)=t_k}}n_i^{t_k}k\frac{n_N^{f(c_i)}}{n^{f(c_i)f(c_N)}}\log\sigma(-V_i\cdot W_N).\end{aligned}\tag{11}$$

As each code c_N maps to code type $f(c_N)$, (11) can be written as

$$\begin{aligned}&\sum_{c_i\in C}\sum_{c_N\in C}n_i^{f(c_N)}k\frac{n_N^{f(c_i)}}{n^{f(c_i)f(c_N)}}\log\sigma(-V_i\cdot W_N)\\&=\sum_{c_i\in C}\sum_{c_j\in C}k\frac{n_i^{f(c_j)}n_j^{f(c_i)}}{n^{f(c_i)f(c_j)}}\log\sigma(-V_i\cdot W_j).\end{aligned}\tag{12}$$

Combining (9) and (12), the objective of typeSkip-gram is

$$\begin{aligned}l=&\sum_{c_i\in C}\sum_{c_j\in C}n_{ij}\log\sigma(V_i\cdot W_j)\\&+\sum_{c_i\in C}\sum_{c_j\in C}k\frac{n_i^{f(c_j)}n_j^{f(c_i)}}{n^{f(c_i)f(c_j)}}\log\sigma(-V_i\cdot W_j).\end{aligned}\tag{13}$$

Let us denote the objective for the code pair c_i and c_j

$$\begin{aligned}l(c_i,c_j)=&n_{ij}\log\sigma(V_i\cdot W_j)\\&+k\frac{n_i^{f(c_j)}n_j^{f(c_i)}}{n^{f(c_i)f(c_j)}}\log\sigma(-V_i\cdot W_j).\end{aligned}\tag{14}$$

To optimize this, we define $x=V_i\cdot W_j$ and calculate the partial derivative with respect to x as

$$\frac{\partial l}{\partial x}=n_{ij}\sigma(-x)-k\frac{n_i^{f(c_j)}n_j^{f(c_i)}}{n^{f(c_i)f(c_j)}}\sigma(x).\tag{15}$$

The partial derivative equals 0 when the local objective reaches its optimum,

$$e^{2x}-\left(\frac{n_{ij}}{k\frac{n_i^{f(c_j)}n_j^{f(c_i)}}{n^{f(c_i)f(c_j)}}}-1\right)e^x-\frac{n_{ij}}{k\frac{n_i^{f(c_j)}n_j^{f(c_i)}}{n^{f(c_i)f(c_j)}}}=0.\tag{16}$$

The solution is $e^x=\frac{n_{ij}}{k\frac{n_i^{f(c_j)}n_j^{f(c_i)}}{n^{f(c_i)f(c_j)}}}$, and

$$\begin{aligned}x=V_i\cdot W_j=&\log\frac{n_{ij}}{k\frac{n_i^{f(c_j)}n_j^{f(c_i)}}{n^{f(c_i)f(c_j)}}}=\log\frac{n_{ij}n^{f(c_i)f(c_j)}}{n_i^{f(c_j)}n_j^{f(c_i)}}-\log k\\=&typePMI_{ij}-\log k,\end{aligned}\tag{17}$$

which is $typePMI$ between c_i and c_j minus constant $\log k$. □

Since typePMI is better at characterizing co-occurrence between codes, the resulting code vectors of typeSkip-gram model should be better than Skip-gram of (3) at capturing relationships between the codes.

3 Experiments

3.1 Dataset

Medical claims used in our experiments come from SEER-Medicare Linked Database [Warren *et al.*, 2002]. In particular, our data contains over 9 million inpatient, outpatient and carrier claims from 161,366 Medicare members diagnosed with breast cancer from 2000 to 2010. From each claim we extracted its set of ICD-9 and CPT codes and disregarded any other information. Our dataset contains 13,977 unique medical codes, including 7,291 ICD-9 diagnosis, 962 ICD-9 procedure and 5,624 CPT codes.

3.2 Experimental Design

Given our dataset of over 9 million medical claims, we treated each claim as a document and for any code occurring in the claim we assumed that all other codes in the claim represent its context. We used several methods to learn vector representations of ICD-9 and CPT codes. As the baselines, we used:

- **SVD**: We applied SVD on standard PPMI matrix which ignores types of medical codes.

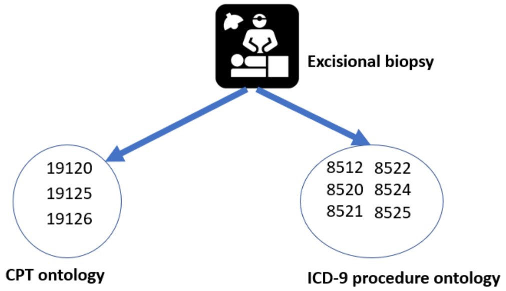

Figure 2: A medical treatment can be encoded with medical codes from different ontologies.

- **Skip-gram**: Using each record as a document, we used the Skip-gram algorithm with negative sampling, which samples negatives using probabilities that are insensitive to the type of medical codes.
- **CBOW**: We also used the continuous bag of words algorithm [Mikolov *et al.*, 2013] which is closely related to Skip-gram.
- **Glove**: Word representations algorithm from [Pennington *et al.*, 2014] is trained on a logarithmic co-occurrence matrix.

We compared the above four baselines with our two proposed approaches, **typeSVD** and **typeSkip-gram**.

The word2vec models (typeSkip-gram, Skip-gram, CBOW) rely on the gradient descent algorithm. In all the experiments we used 60 iterations, as we empirically observed that it is sufficient for the vector representation to stabilize. The number of negative samples was set to 5.

As shown in Figure 2, ICD-9 procedure and CPT codes both describe medical procedures performed by healthcare providers, while CPT has higher granularity. Clinical researchers rely heavily on both ICD-9 procedure and CPT codes to identify treatment and define cohorts from massive clinical data [Bleicher *et al.*, 2016; Bleicher *et al.*, 2012]. However, it is time-consuming and burdensome to identify the same concepts in different ontologies. For example, physicians are more familiar with CPT codes and can easily find that medical procedure "mastectomy" is encoded as CPT code "19180," while they might have difficulty finding ICD-9 procedure codes corresponding to the same concept. Therefore, there is a need to create translations between different ontologies. Currently, there is a lack of a reliable official mapping between different ontologies.

To evaluate whether our proposed multi-source code representation algorithms could be helpful in finding correspondence between ICD-9 procedure codes and CPT codes, in this paper we used the gold standard mappings for procedures related to breast cancer manually derived by clinical researchers [Bleicher *et al.*, 2012]. This mapping recognizes 18 procedures commonly used to treat breast cancer such as excisional biopsy, mastectomy, and breast MRI. Clinical researchers rely on these codes to identify patients diagnosed with breast cancer. For each procedure, a list of the corresponding CPT and ICD-9 procedure codes is provided. A total of 149 CPT codes and 58 ICD-9 procedure codes are listed (the details are provided in the Appendix[2]). For example, concept group 17 "brain MRI" contains CPT codes "70551"-"70553" and ICD-9 procedure code "8891". Good vector representation should place closely related ICD-9 procedure codes and CPT codes in the vicinity. Therefore, given ICD-9 code "8891", users should be able to easily find the CPT counterparts by nearest neighbor search in the embedding space.

Since retrieving related codes among different ontologies is an information retrieval task, we used Normalized Discounted Cumulative Gain (NDCG) [Järvelin and Kekäläinen, 2002], a measure of ranking quality commonly used in information retrieval, to evaluate the quality of the resulting vector representations. For each of the 18 concept groups from our gold standard we calculated two NDCG scores in the following way. Let us assume that k-th concept group includes n ICD-9 procedure codes $\{icd_1, icd_2, ..., icd_n\}$ and m CPT codes $\{cpt_1, cpt_2, ..., cpt_m\}$.

Task 1 (CPT → ICD-9 procedure): For each CPT code cpt_i in concept group k, find its p nearest ICD-9 procedure codes based on the cosine similarity. ICD-9 code ranked at position q is assigned label $r_q = 1$ if it is from the same concept group k and $r_q = 0$ otherwise. The $NDCG_p$ score of the CPT code cpt_i is calculated as

$$\begin{aligned} DCG_p &= \sum_{i=1}^{p} \frac{r_i}{log(i+1)} \\ IDCG_p &= \sum_{i=1}^{n} \frac{1}{log(i+1)} \\ NDCG_p &= \frac{DCG_p}{IDCG_p}. \end{aligned} \tag{18}$$

Assuming $p > n$, $ICDG$ represents the upper bound on the value of DCG. $NDCG$ score ranges from 0 to 1. The higher the score, the better the ranking quality. For each concept group, we report the average $NDCG_p$ of all m CPT codes in that group.

Task 2 (ICD-9 procedure → CPT): Similarly to Task 1, for each ICD-9 procedure code icd_i in concept group k, find its p nearest CPT codes. We report the average $NDCG_p$ of all ICD-9 procedure codes for each of the 18 concept groups.

3.3 Results

In Table 1 and Table 2, we show the summary results for Tasks 1 and 2. Each entry in the tables is an average $NDCG$ over all 18 concept groups for a given algorithm and a given choice of p. We report $NDCG_{20}$, $NDCG_{50}$ and $NDCG_{100}$ values. The detailed results for each of the 18 groups are provided in the Appendix. The main observation is that the proposed multi-source algorithms typeSVD and typeSkip-gram are superior to the baselines on both tasks. This is a strong indicator that paying attention to the types of codes can result in improved code representations for cross-referencing of medical codes. While the observed difference between typeSVD

[2] https://github.com/AU19/IJCAI19/blob/master/Appendix.pdf

	p=20	p=50	p=100
SVD	0.54	0.55	0.55
Skip-gram	0.619	0.621	0.622
CBOW	0.541	0.555	0.56
Glove	0.622	0.612	0.608
typeSVD	0.64	0.638	0.637
typeSkip-gram	**0.642**	**0.641**	**0.644**

Table 1: Average $NDCG_p$ of all concept groups on Task 1

	p=20	p=50	p=100
SVD	0.391	0.442	0.459
Skip-gram	0.504	0.51	0.536
CBOW	0.365	0.398	0.416
Glove	0.443	0.456	0.454
typeSVD	**0.582**	0.569	**0.588**
typeSkip-gram	0.557	**0.585**	0.576

Table 2: Average $NDCG_p$ of all concept groups on Task 2

ICD-9 Codes	Concept Group	Description
85.21	1	Local excision of lesion of breast
85.19	6	Other diagnostic procedures on breast
85.22	1	Resection of quadrant of breast
40.11	5	Biopsy of lymphatic structure
40.23	5	Excision of axillary lymph node
85.23	2	Subtotal mastectomy
40.19	6	Other diagnostic procedures on lymphatic structures
40.3	5	Regional lymph node excision
40.51	5	Radical excision of axillary lymph nodes
85.41	3	Unilateral simple mastectomy
85.43	4	Unilateral extended simple mastectomy
38.52	NA	Ligation and stripping of varicose veins
85.12	1	Open biopsy of breast
85.20	1	Excision or destruction of breast tissue not otherwise specified
92.16	6	Scan of lymphatic system
86.3	NA	Other local excision or destruction of lesion or tissue of skin and subcutaneous tissue
85.24	1	Excision of ectopic breast tissue
40.29	5	Simple excision of other lymphatic structure
40.22	5	Excision of internal mammary lymph node
19.12	NA	stapedectomy

Table 3: Nearest 20 ICD-9 procedure codes of CPT code “19120” (Removal of breast lesion), ranked by cosine similarity.

and typeSkip-gram is rather small, it is interesting that Skip-gram is much more accurate than SVD. The large improvement between SVD and typeSVD is an indicator that SVD is much more vulnerable to the inaccurate PMI values than is the Skip-gram to the choice of negative sampling probabilities. There is a relatively small difference in the relative performance of the algorithms as a function of p.

To gain a further insight into the performance of our algorithms on the cross-referencing of medical codes, in table 3 we list the nearest 20 ICD-9 procedure codes of CPT code “19120” described as “removal of breast lesion” and which is assigned to the “excisional biopsy” concept group in the gold standard. The ranking is based on vector representations obtained from typeSkip-gram algorithm.

It could be observed that five ICD-9 procedure codes from the same concept group “excisional biopsy” are among the 20 nearest ICD-9 procedure codes of CPT code “19120”. The only missing ICD-9 code “85.25” was ranked as 65th. The $NDCG_{20}$ score calculated using (18) for CPT code “19120” is 0.77. It is interesting to observe that among the 20 nearest neighbors, 12 belong to other concept groups from the gold standard and only 3 are not part of the gold standard. Those 12 related codes are also related to breast cancer and their presence in the list indicates that they occur in the similar context as CPT code “19120”. This result hints that a good approach for cross-referencing of codes would be to produce a list of neighbors and have human experts focus on those codes rather than having to make guesses or sift through a much larger set of candidate codes.

Of the 3 ICD-9 codes not listed in the gold standard, two of them, “38.52” and “86.3”, are also closely related to breast cancer related surgeries. However, code “19.12” “Stapedectomy” is related to a surgical procedure of the middle ear in order to improve hearing. It clearly looks like an outlier. Interestingly, after a follow-up study of claims that contain both this code and codes describing excisional biopsy, it became apparent that this is a consequence of coding entry errors. In particular, it seems that the coders are occasionally trying to enter CPT code “19120” into the ICD-9 procedure field which allows only 4 numbers. Thus, they would enter only the first 4 numbers (i.e., 1912) and this would end up being recorded as a middle ear surgery. This interesting insight is another proof that vector representations by the typeSkip-gram algorithm are of high quality and it also hints at another possible application for discovery of coding errors.

4 Conclusion

In this paper, we have proposed a new approach to learn vector representations of medical codes from medical claims coming from different types of providers. Our first contribution was to propose a modification to the Pointwise Mutual Information (PMI) measure between the codes and our second contribution was to propose a new negative sampling method for word2vec model that is implicitly factorizing the modified PMI matrix. The new approach is evaluated on the cross-referencing between ICD-9 and CPT coding ontologies using a gold standard expert mapping related to breast cancer. Our results indicate that vector representations of codes learned by the proposed approach outperform the baselines. Future work can use the proposed method on the problem of cross-referencing of ICD-9 and ICD-10 codes. More generally, the proposed method is applicable to embedding of items contained in heterogeneous types of data sets, where the distribution of items being embedded differs among the sources.

Acknowledgments

This was funded by NIH/NCI grants R03CA152388, R21CA202130, and P30CA006927.

References

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

[Bai *et al.*, 2017] Tian Bai, Ashis Kumar Chanda, Brian L Egleston, and Slobodan Vucetic. Joint learning of representations of medical concepts and words from ehr data. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 764–769. IEEE, 2017.

[Bai *et al.*, 2018a] Tian Bai, Ashis Kumar Chanda, Brian L Egleston, and Slobodan Vucetic. Ehr phenotyping via jointly embedding medical concepts and words into a unified vector space. *BMC medical informatics and decision making*, 18(4):123, 2018.

[Bai *et al.*, 2018b] Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 43–51. ACM, 2018.

[Bleicher *et al.*, 2012] RJ Bleicher, K Ruth, ER Sigurdson, E Ross, YN Wong, SA Patel, M Boraas, NS Topham, and BL Egleston. Preoperative delays in the us medicare population with breast cancer. *Journal of Clinical Oncology*, 30(36):4485–4492, 2012.

[Bleicher *et al.*, 2016] Richard J Bleicher, Karen Ruth, Elin R Sigurdson, J Robert Beck, Eric Ross, Yu-Ning Wong, Sameer A Patel, Marcia Boraas, Eric I Chang, Neal S Topham, et al. Time to surgery and breast cancer survival in the united states. *JAMA oncology*, 2(3):330–339, 2016.

[Brouch, 2004] K Brouch. Ahima project offers insights into snomed, icd-9-cm mapping process. *Health Information Management*, 32(1):31–34, 2004.

[Butler, 2007] RR Butler. Icd-10 general equivalence mappings: Bridging the translation gap from icd-9. *Journal of AHIMA*, 78(9):84–86, 2007.

[Cai and Wang, 2018] Liwei Cai and William Yang Wang. Kbgan: Adversarial learning for knowledge graph embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, USA, 2018. ACL.

[Cai *et al.*, 2018] Xiangrui Cai, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, Ying Zhang, and Xiaojie Yuan. Medical concept embedding with time-aware attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3984–3990. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[Choi *et al.*, 2016a] E Choi, MT Bahadori, E Searles, C Coffey, and J Sun. Multi-layer representation learning for medical concepts. In *KDD*, 2016.

[Choi *et al.*, 2016b] Y Choi, CY Chiu, and D Sontag. Learning low-dimensional representations of medical concepts. In *AMIA Summits on Translational Science Proceedings*, pages 123–345, 2016.

[Choi *et al.*, 2017] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM, 2017.

[Grbovic and Cheng, 2018] Mihajlo Grbovic and Haibin Cheng. Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &#38; Data Mining*, KDD '18, pages 311–320, New York, NY, USA, 2018. ACM.

[Järvelin and Kekäläinen, 2002] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[Levy and Goldberg, 2014] O Levy and Y Goldberg. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 2177–2185, 2014.

[Mikolov *et al.*, 2013] T Mikolov, I Sutskever, K Chen, G Corrado, and J Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, 2013.

[Pennington *et al.*, 2014] J Pennington, R Socher, and CD Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[Schulz *et al.*, 1998] S Schulz, A Zaiss, R Brunner, D Spinner, and R Klar. Conversion problems concerning automated mapping from icd-10 to icd-9. *Methods of Information in Medicine*, 37(3):254–259, 1998.

[Topaz and Shafran-Topaz, 2013] M Topaz and L Shafran-Topaz. Icd-9 to icd-10: evolution, revolution, and current debates in the united states. *Perspectives in Health Information Management*, page 1, 2013.

[Turney and Pantel, 2010] PD Turney and P Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37:141–188, 2010.

[Wang *et al.*, 2018] Peifeng Wang, Shuangyin Li, and Rong Pan. Incorporating gan for negative sampling in knowledge representation learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Warren *et al.*, 2002] Joan L Warren, Carrie N Klabunde, Deborah Schrag, Peter B Bach, and Gerald F Riley. Overview of the seer-medicare data: content, research applications, and generalizability to the united states elderly population. *Medical care*, pages IV3–IV18, 2002.