Topic Medical Concept Embedding: Multi-Sense Representation Learning for Medical Concept

1st Feng Qian[1] *, 1st Chengyue Gong[3] *, Lu-chen Liu[1], Lei Sha[2], Ming Zhang[1] †
[1] Institute of Network Computing and Information Systems
School of Electronics Engineering and Computer Science, Peking University
[2] Key Laboratory of Computational Linguistics, Ministry of Education
School of Electronics Engineering and Computer Science, Peking University
[3] Department of Information Management, Peking University
{nickqian, cygong, liuluchen292, shalei, mzhang_cs}@pku.edu.cn

***Abstract*—Representation learning algorithm in medical area maps high dimensional real world medical concepts to low dimensional vector space, encodes rich medical knowledge, and has brought improvement to various machine learning applications in medical area. However, previous representation learning models in medical area failed to consider the multi-sense characteristic of medical concept. Moreover, the inner relationships between representations learned by previous model is implicit and can only be explained according to visualization, which means poor interpretability. In this paper, we propose Topic Medical Concept Embedding (TMCE), a generative embedding model to address above two problems. TMCE is able to learn multi-sense representations for a single medical concept, and TMCE can also improve interpretability by modeling relationships between each concept explicitly. In TMCE, multi-sense concept representations are influenced by its contexts and its topics. In addition, dosage information which is ignored by previous work are also utilized in TMCE. A MCMC method is presented to jointly learn the two-layer topic embeddings and multi-sense concept embeddings. Experimental results show that representations learned by TMCE outperforms those learned by other strong baselines by a large margin in a multi-label diagnose classification tasks. Several case studies further show that TMCE can learn medically correct multi-sense representations with better interpretability than other strong baselines.**

***Index Terms*—representation, multi-sense, topic modeling, context, interpretability**

I. Introduction

Learning low-dimensional continuous representations for discrete high dimensional medical concepts has attracted a lot of attention. These algorithms transform high dimensional real world medical concepts into lower dimensional real value vectors while still keeping inner relationships, for example, medicines with similar usage will be close in the lower dimensional space. Such representations can help capture latent knowledge behind medical concepts and have shown improvement on various machine learning applications [1]–[3]. Although representations learned by previous proposed work have brought improvements over accuracy of various prediction tasks, they all failed to address two important problems:

- One medical concept may have more than one complete different meanings, for example, aspirin can be used to cure fever as well as cardiovascular disease. Existing methods can only assign one representation for each medical concept, which is not suitable in many cases. In addition, note the fact that in electronic health record (EHR), medical concepts are always recorded with dosage information which is numerical value corresponding to medical concepts. Note that change in dosage information can change the implication of a medical concept completely, so dosage is informative to find different meanings of a same concept. For example, a heart rate at 180 beats per minute (bmp) means danger while a heart rate at 90 bpm only means normality. None of previous work is able to utilize this useful information.
- A requirement of machine learning methods in medical area is strong interpretability. Relationships between representations learned by previous methods is implicit, that is to say relationships between representations can only be explained according to visualization empirically. An algorithm that can explicitly and automatically model the inner relationships between different representations and provide better interpretability is needed.

In this paper, we propose Topic Medical Concept Embedding (TMCE) for medical concept to address the above two problems. TMCE combines intuitions from both embedding methods that exploit context, and non-parametric topic models that model topic structure. Embedding methods learn representations by analyzing context around each data point. These methods can learn from context, however, relationships between learned representations are implicit and can only be explained empirically according to visualization. While, non-parametric topic models can group data into topics, and automatically derive explicit explanation according to learned topics. In TMCE, both intuitions are combined, and multi-sense representations are learned by utilizing both context information and topic structure. In addition, dosage information are also utilized by TMCE to enhance performance.

The architecture of TMCE is shown in Figure 1. As shown

* These authors contribute same to this work
† Corresponding Author

in Figure 1, a two-layer Dirichlet processes is built to model implicit relationships behind representations. The global layer Dirichlet process in layer C models all possible medical topics that are shared by different medical concepts in local layer K. Each local layer Dirichlet process in a concept models related topics for current medical concept, under the influence of dosage information. Multi-sense representations of a concept are drawn from corresponding local topics under the influence of context information. In this way, representations in TMCE are influenced by both the local topics which they belong to, and context information around them. TMCE is trained with MCMC algorithm.

Experimental results show that representations learned by TMCE outperform other strong baselines including skip gram and stack AutoEncoder by large margin on a multi-label diagnose classification task. In addition, several experiments using visualization and other methods are given to show that multi-sense representations learned by TMCE are medically correct and hold better interpretability.

II. Related Works

The popularity of representation learning started from Natural Language Process community. Inspired by the success representation learning gained in NLP, representation learning also attracted a lot attention in medical machine learning field. Previous representation learning methods in medical area can be classified into two categories: The first category includes works that learn representations from free text. Minarro et al. applied skip gram algorithm directly to medical text corpus collected from websites such as PubMed to learned medical concepts representation [4]. De et al. learned concept representations from patient text records and medical journal abstract after replacing words in the corpus by Unified Medical Language System (UMLS) standard concept vocabulary [5]. The second category includes works learn representations from electronic health record (EHR), which is structured medical data collected by hospitals. Choi et al. learned concept representations with skip gram on EHR data [2], [3]. In order to learn representation for a single visit, Choi et al. used a neural network and skip-gram-like objective to learn representations for both each medical concept and each visit made by a patient [1]. However, previously proposed methods failed to consider the multi-sense characteristic of medical concepts, and relationships between learned representations are implicit, so the interpretability is limited.

A. Dirichlet Process

Dirichlet process (DP) was first introduced by Ferguson et al. [6]. DP is characterized by an arbitrary base distribution G_0 and a positive concentration parameter α, and is a probability distribution whose range is itself a set of probability distributions. DP is often used in Bayesian inference to describe the prior knowledge about the distribution of random variables. A distribution G can be drawn from DP as shown in Equation 1, and M random variables $\{\boldsymbol{\theta}_m\}$ are drawn from G as shown in Equation 2:

$$G|\alpha, G_0 \sim DP(\alpha, G) \tag{1}$$

$$\boldsymbol{\theta}_m|G \sim G, \quad m = 1, ...M \tag{2}$$

The Dirichlet process can be seen as the infinite-dimensional generalization of the Dirichlet distribution. As Dirichlet distribution is the conjugate prior for the categorical distribution, the Dirichlet process is the conjugate prior for infinite, non-parametric discrete distributions. In this perspective, DP can be used as a prior on the parameters of a mixture model of unknown numbers of components. For example, given the first $M-1$ samples of G, the new sample $\boldsymbol{\theta}_M$ is either drawn from existing draws, according to a multinomial allocation based on the number of previous draws, or drawn from the base distribution G_0 with probability $\frac{\alpha}{\alpha+M-1}$.

III. Multi-Sense Representation Learning With Amount For Medical Concept

In this section, first we give clear definitions to notions that are used in this work. Then, we will give details on architecture of TMCE.

- **Medical Concept:** Medical concept in our work includes medication (such as aspirin) and medical indicators (such as heart rate). Medical concepts are denoted as c_i.
- **Dosage:** In data, regarding one concept, only a certain number of dosages have appeared. m_{ij} is the j^{th} possible dosage associated with concept c_i in data.
- **Context-dosage pair:** Since every concept has a dosage, for simplicity and clearer notation, we call a concept with dosage (c_i, m_{ij}) concept-dosage pair.
- **Context:** The context for a concept-dosage pair means all other concepts-dosage pairs that are recorded in a same visit made by a patient.
- **Concept representation:** Each medical concept will be assigned several multi-sense representations, which are called multi-sense representations. c_{ij} is the indicator of which representation is allocated (c_i, m_{ij}), and the corresponding representation is marked as v_{ij}.

Our multi-sense embedding model is designed under such an assumption that each medical concept can be related to various topics, and different medical concepts can share a same topic. Under this assumption, we designed an embedding model using a two-layer Dirichlet process to model implicit relationships between representations explicitly. Multi-sense representations of one certain concept are drawn under the influence of related topics and context. Dosage information are also taken into account. The architecture of our model is shown in Figure 1.

A. Global Layer of Topic Allocation

In global layer of TMCE, which is shown as layer C in Figure 1, various medical topics are modeled. Note the fact that there are many different types of topics in medical area, and the number of these topics is unknown at the start of modeling, it is natural to model the mixture of all the topics by a Dirichlet process, which is a non-parametric prior for

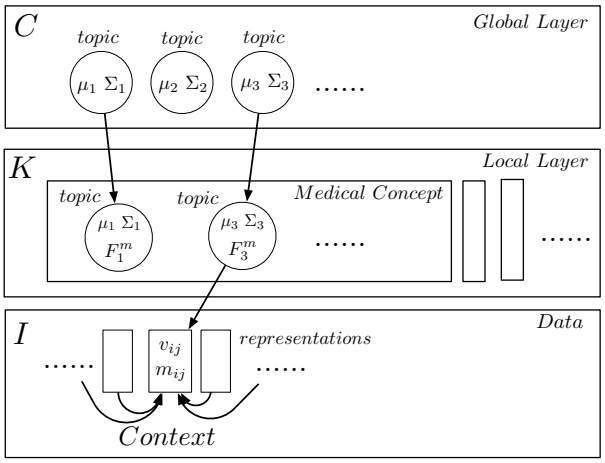

Fig. 1: Architecture of TMCE

a mixture with unknown numbers of component. A global Dirichlet process is designed, with a base distribution $H = NIW(\mu_0, \Sigma_0, \nu, \lambda)$ representing all possible topics of medical concept:

$$\begin{aligned} G_0|\gamma, H &\sim DP(\gamma, H) \\ (\mu_c, \Sigma_c) &\sim G_0 \end{aligned} \tag{3}$$

$|C|$ Normal distributions, each represents one topic, are drawn from the distribution G_0.

B. Local Layer of Topic Allocation

Global layer of TMCE models all possible topics, and these topics are shared by medical concepts in local layer, which is shown as layer K in Figure 1. Local layer is comprised of medical concepts, and each medical concept can be allocated one or more topics from global layers. As explained before, dosage information can provide valuable information, so local dosage distributions are mixed with global topic distributions to from local topic distributions. To be more specific, for each medical concept i, the allocation G_i for each representation v_{ij} is drawn from a local Dirichlet process.

$$\begin{aligned} G_i' &= (G_0, F_i^M) \\ G_i|\alpha, G_i' &\sim DP(\alpha, G_i') \end{aligned} \tag{4}$$

where G_i' is made up of global measure of topics G_0 and local measure of dosage F_i^M, under the assumption that dosage information can be used as auxiliary evidence for topics and embedding inference.

As shown in Figure 1, each concept in local layer K is made up of pairs of global measurement of topic and local measurement of dosage. Topics are shared among concepts, and a concept may be related to multiple topics. When a concept is associated with a specific kind of topic, the dosage should follow a distribution with a local dosage measurement as parameter. A intuitive explanation is: aspirin (a concept) may have different usages (local topic allocations), when aspirin is used to cure fever (a specific local topic), corresponding dosage should be around 50mg (local dosage measurement).

C. Draw Representation

Finally, both information from the two level topic embedding structure, and information from context are combined to generate representation for c_{ij} as shown in Equation 5.

$$v_{ij} \sim p(v_{ij}|\mu_j, \Sigma_j, \boldsymbol{v}_{ij,context}) \tag{5}$$

Dosage is drawn from a Normal distribution and the local measure is $F_i^m = N(\mu_i^M, \Sigma_i^M)$:

$$m_{ij} \sim N(m_{ij}|\mu_{ij}^M, I) \tag{6}$$

In our model, global layer DP generates parameters μ_j and Σ_j for each topic according to the base distribution H which is a Normal-Inverse-Wishart distribution and concentration parameter η. Then, the scaling parameter ν adjusts between cluster scattering of the cluster-specific mean vectors. These parameters, together with different base measures of the dosage F_i of each medical concept i, define the base measures G_i' of the local layer DPs. Finally, representations and corresponding dosage are drawn from local DP under the influence of context.

IV. Inference

In this section, posterior inference is used to sample parameters from their joint posterior distribution given observed data. To simplify denotation, we use indicator to denote which topic does the current unit derive from. z_{ij} is used as the indicator of (v_{ij}, m_{ij}) in the local DP layer, and t_{ik} is used as the indicator of k component of the medical concept c_i in the global DP layer.

Marginalizing out the base distribution, z_{ij} and t_{ij} can be sample in:

$$\begin{aligned} &p(z_{ij} = k|\boldsymbol{V}, \boldsymbol{m}, \boldsymbol{z}_i^{-j}, \boldsymbol{t}) \\ &\propto \begin{cases} \alpha p(v_{ij}) p(m_{ij}) & k = K+1 \\ n_i^{-j} N(v_{ij}|\mu_k^{-j}, \Sigma_k^{-j}) N(m_{ij}|\mu_k^M) & k \leq K \end{cases} \end{aligned} \tag{7}$$

$$\begin{aligned} &p(t_{ik} = c|\boldsymbol{V}, \boldsymbol{m}, \boldsymbol{t}_i^{-k}, \boldsymbol{z}) \\ &\propto \begin{cases} \gamma \prod_{j:z_{ij}=k} p(v_{ij}) & c = C+1 \\ o_c^{-k} \prod_{j:z_{ij}=k} p(v_{ij}|\mu_i^{-k}, \Sigma_i^{-k}) & c \leq C \end{cases} \end{aligned} \tag{8}$$

where o_c and n_i represent the number of samples of the global and local topics respectively, V stands for all representations, m stands for all dosage in data. While sampling the t_{ik}, we only update global topic parameters and assume that removing a single data instance does not significantly change local topic parameters. Because of the dependency among data instances, removing a data instance from a global topic not only affect the parameters of the global topic it belongs to, but also the corresponding local topic parameters. However, updating local topic parameters for every data instance removed will be too computational costly, so the influence on local topic parameters is ignored for the sake of computational efficiency.

Then, the parameters of the local and global topics are sampled. The posterior distribution of the component parameters can be obtained by marginalizing out the joint distribution

$$\begin{aligned} p(\mu_c|\boldsymbol{z},\boldsymbol{t},\boldsymbol{V},\Sigma_c) &\sim N(\frac{\lambda\mu_0+n_c\bar{\boldsymbol{v}}}{\lambda+n_c},\frac{\Sigma_c}{\lambda+n_c}) \\ p(\Sigma_c|\boldsymbol{z},\boldsymbol{t},\boldsymbol{V},\mu_0,\Sigma_0) &\sim IW(\nu+n_c,\Sigma_c') \\ p(\mu_{ik}^M|\boldsymbol{m}_i) &\sim N(\frac{\lambda\mu_0+n_k\bar{\boldsymbol{v}}}{\lambda+n_k},\frac{\Sigma_i^M}{\lambda+n_k}) \end{aligned} \tag{9}$$

where $\Sigma_c' = \Sigma_0 + \boldsymbol{S} + \frac{\lambda n_c}{\lambda+n_c}(\bar{\boldsymbol{v}}-\boldsymbol{\mu_0})^{\boldsymbol{T}}(\bar{\boldsymbol{v}}-\boldsymbol{\mu_0})$ and $\boldsymbol{S}$ is the scatter matrix.

Given all the other parameters, the posterior of embedding vectors are drawn from the conditional posterior distribution $p(v_{ij}|z_{ij}, v_{ij,context})$ described in Equation 5. In order to accelerate the training, the correlation between context information and topic information is ignored, that is to say we assume they are conditionally independent given the embedding v_{ij}, so that the format of the objective can be simplified into:

$$\max_{v}\{\prod(\prod_{context} p(v_{ij,context}|v_{ij}))p(v_{ij}|z_{ij})\} \tag{10}$$

In this way, the learning process is divided into two stages: In the first stage, gradient optimization is used on the conditional distribution of V, which is in the format of skip-gram. In the second stage, V is treated as fixed and Bayeisan inference is taken place to find the optimal variables of our model.

V. Experiment

We evaluate the performance of TMCE on MIMIC3[1] dataset in both quantitative and qualitative method. First we give detailed description about the data used to train and evaluation TMCE. Then we compare TMCE with other popular baselines on a diagnose multi-label classification task. After that, more experiments are presented to show TMCE can bring about multi-sense feature and better interpretability. These experiments show the interpretability of concept multi-sense representations, local topic and global topic allocation.

A. Settings

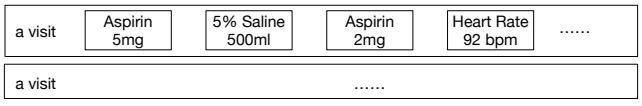

Fig. 2: A Example of Data Structure in MIMC3

We evaluate performance of TMCE on a public available dataset MIMIC3 [1], which is a deidentified health dataset. A piece of sample from the dataset are shown in Figure 2. Each time a patient pays a visit, all data including medication usage, medical body indicator, medical test result, diagnoses from doctor etc. are recorded during the visit. More specifically, we use data in following tables in MIMIC3: "CHARTEVENTS", "INPUTEVENTS_CV", "INPUTEVENTS_MV", "LABEVENTS", "MICROBIOLOGYEVENTS". For computational efficiency, we only choose to model the most frequent 1000 medical concepts in MIMIC data. Statistical information of data we used are: number of visit: 58077; number of unique medical concept: 928; average number of unique medical concepts per visit: 210; number of unique concept dosage pairs: 333613; average number of unique medical concept per visit: 884.

In the table "DIAGNOSE_ICD" in MIMIC3, each visit was assigned several International Classification of Disease (ICD) codes for billing and recording purpose. ICD codes are diagnoses given by doctors to each patient during a single visit. To preform diagnoses prediction evaluation, we extract ICD codes for each visit, and assign the ICD codes as tags to every concept-dosage pair within the visit. Note that a same concept-dosage pair may appear in different visits, a concept-dosage pair may receive ICD tags from different visits.

B. Multi-label classification

For each diagnose d_i, the objective is to predict which concept-dosage pair is associated with this diagnose. More specifically, for each d_i, classify all concept-dosage pairs into two classes according to whether this concept-dosage pair has d_i as tag. The multi-sense concept representations are input to a one-vs-rest logistic regression classifier with L2 regularization. Logistic regression classifiers are trained on 10%-90% data respectively and tested on the rest of the data and ten-fold cross validation are applied. For all diagnoses involved in the prediction task, $F\ Score$ is calculated and shown in Table I. We compare TMCE with following strong baselines:

- **Skip Gram without dosage** When training skip gram without amount, we simply ignore dosage information, and regard all c_{ij} as c_i. When testing, we use representation of c_i for all c_{ij}
- **Skip Gram with dosage** Apply skip gram to every concept with different dosage in data. Notice that by skip gram, only context information can be captured.
- **Stack AutoEncoder** Stacked AutoEncoder is a popular unsupervised representation algorithm proposed by Vincent et al. [7]. We trained a 3-layer (256-128-64) stacked autoencoder from randomly initialized one-hot representation.

Since one-hot representation will be much too high dimensional, and one-hot representation can not provide any interpretability at all, we do not set one-hot as one of our baselines.

In our model, the hyper-parameters are set as: $\alpha = 0.01, \eta = 0.02, \lambda = 2, \Sigma = diag\{1, 1, ..., 1\}$, and ν is set to one adding the dimension of embedding vector.

Experimental results in Table I show that TMCE outperforms all other baselines in multi-label diagnose classification task by a large margin. There are two main reasons why TMCE can outperform all other baselines: the using of the dosage

[1]https://mimic.physionet.org/

TABLE I: Experimental result of multi-label diagnose classification task. Percentage on the top stands for the percentage of all data used as training data.

		10%	20%	30%	40%	50%	60%	70%	80%	90%
F Score	Skip-Gram without dosage	0.673	0.673	0.680	0.692	0.692	0.697	0.710	0.718	0.712
	Skip-Gram with dosage	0.670	0.673	0.676	0.680	0.685	0.693	0.703	0.718	0.739
	SAE	0.656	0.651	0.654	0.669	0.674	0.681	0.703	0.713	0.737
	TMCE	**0.686**	**0.695**	**0.709**	**0.723**	**0.738**	**0.755**	**0.770**	**0.785**	**0.798**

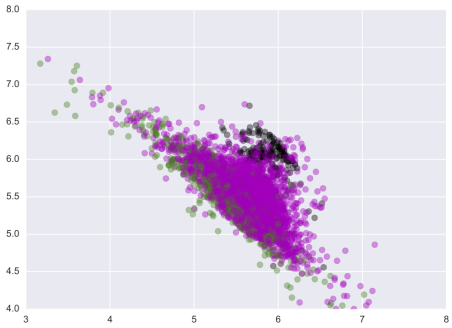

Fig. 3: 1st Training Epoch by TMCE

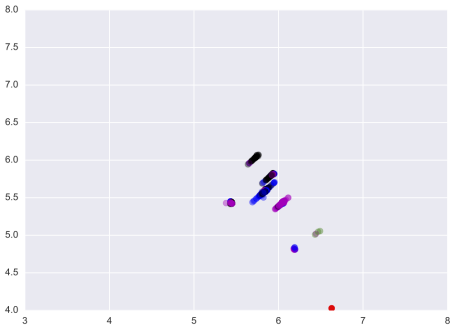

Fig. 4: 2st Training Epoch by TMCE

Fig. 5: 3st Training Epoch by TMCE

information and topic information. First, in contrast to skip gram without amount, TMCE utilizes extra dosage information which can change the implication of a concept completely in some situations, for example, it is apparent that diagnoses associate with a high heart rate should not be the same as those associated with a moderate or low heart rate. To model and train representations on a concept-dosage granularity will help encode richer information, and brings absolute advantage.

Second, topic information utilized by TMCE helps TMCE outperform skip gram with dosage. According to Equation 5, the topic information is added into the objective, and that is to say, representations of concept-dosage pairs with the same topic will get closer to each other. Representations learned directly from context by skip gram with dosage are deteriorated by stochastic noise in data since in reality the amount of data is not sufficient enough to clear all the noise and generate perfect inference. However, by utilizing common pattern in a group of similar representations, noise can be offset between concept-dosage pair in a same topic.

To better understand the advantage brought by using topic information, we can compare performance of skip-gram with dosage and skip-gram without dosage. Even though skip gram with dosage utilizes extra dosage information than skip gram without dosage, skip gram with dosage performs worse. The reason lies in the extra noise brought by adding dosage information. More stochastic noise are also brought in by adding dosage information because more concept-dosage pairs are to be embedded than concept itself. More concept-dosage pairs to be embedded brings more stochastic noise, thus, the benefit brought by adding more dosage information are offset by extra noise. TMCE solves this problem by using topic information to offset stochastic noise between concept-dosage pairs among a same topic, so that TMCE can outperform two baselines.

The result of the stack AutoEncoder is learned from one-hot vectors, which have neither context information nor topic information. Lack of information encoded is the reason why stackAutoEncoder gives a low performance in multi-label diagnose classification task.

C. Representation Interpretability

Representations learned by TMCE have the capability to be multi-sense and holds better interpretability. To further elaborate the advantage of TMCE, visualization of the training process is shown in Figure 3. Principle component analysis (PCA) algorithm is used to visualize representations in the process of training. In the pictures, all concept-dosage pairs of medical concept Aspirin are shown, where one color stands for a topic. The first picture shows the first epoch of TMCE in Figure 3. Because of the two-level topic allocation structure designed in TMCE, all concept-dosage pairs are divided into a few meaningful topics. However, because of stochastic noise on each concept dosage pair, concept-dosage pairs belongs to different global topics mix together, and different topics overlap with each other and disperse widely in the embedding feature space. In Figure 4, after two epochs of TMCE training, concept-dosage pairs belong to a same global topic get closer. Intuitively, concept-dosage pair with similar implication should have a similar representation. In the third epoch shown in Figure 5, information from concept, topic and dosage are better utilized, and the distance between different topics increases, the boundary between different global topics become clearer.

Since each concept may be allocated with several global topics, and topics are shared among different concepts, one global topic can match multiple concepts. If TMCE can correctly model the inner relationships between different concepts, concepts that have a same topic allocated should be related to each other closely in terms of medical relationships. In Table II, three global topics are listed. Under each topic, concepts that are related to it are listed. We choose three global topics of concepts with moderate concept numbers

TABLE II: Three Sample Global Topics And Related Medical Concepts

Sample Topic 1	**Sample Topic 2**	**Sample Topic 3**
Intubation	Calcium	Insulin-Humalog
Platelet Count	Po Intake	Insulin-Glargine
PO2	Invasive Ventilation	Dextrose 5%
Bicarbonate	O_2 Flow	Insulin-Regular
Blood Cultured	OR Crystalloid	Morphine Sulfate
Red Blood Cells	Fibrinogen	Lorazepam (Ativan)

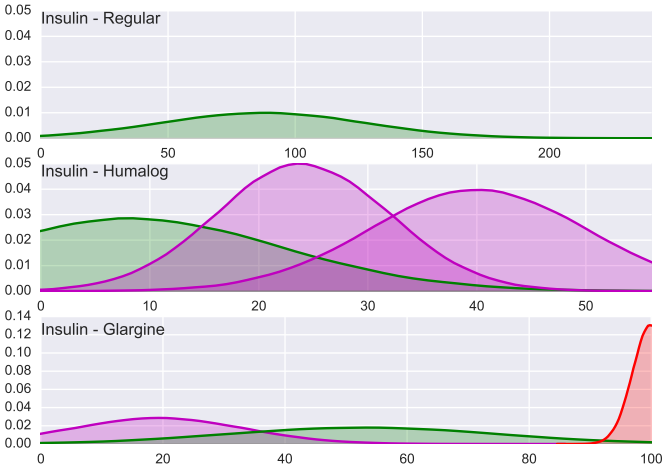

Fig. 6: Local Visualization of Mixture of 3 Different Kinds of Insulins

for the sake of display. In sample topic 1, concepts about blood, blood testing are clustered together. In sample topic 2, concepts related to nutrition and supplementary medication for surgery are clustered together. In sample topic 3, concepts about insulin and anesthetic are clustered together. TMCE has modeled the inner relationships between different concepts correctly, and in contrast to previous methods that explain the inner relationships between concepts according to visualization, TMCE learns the inner relationships automatically and explicitly using topic structure, which brought much better interpretability than previous works.

In Figure 6, local topics of three related medical concepts, insulin-regular, insulin-humalog, and insulin-glargine are shown. Each distribution stands for the dosage distribution of one local topic. Insulin-regular is a kind of regular insulin, insulin-humalog is a kind of fast-acting and stronger insulin, while insulin-glargine is a kind of slow-acting insulin. To make it clear to understand, a same color in three mixtures stands for a same global topic, and for convenience, we will refer to three global topics as topic-green, topic-pink and topic-orange. First, all topics assigned to three types of insulin are among topic-green, topic-pink, and topic-orange, and this demonstrates that TMCE recognizes the similarity between three kinds of insulin, so the topics assigned to this three kinds of insulins are the same. Second, note that smaller dosage of fast rate insulin, medium dosage of regular insulin, and larger dosage of slow rate insulin are assigned the same topic, topic green. Such pattern is medically correct since compares to regular insulin, larger dosage of slow-effect insulin is needed in order to reach the same effect, while only smaller amount of fast-rate insulin is needed. The example shows that TMCE is able to learn multi-sense representations with the help of dosage information. The learned multi-sense representation is medically correct which further validates the ability of TMCE to provide multi-sense representations with better interpretability.

VI. Conclusion

In this paper, we propose TMCE to learn multi-sense representations for medical concept. TMCE is also able to take dosage information into account, and learns the relationships between representations explicitly which brings better interpretability. TMCE combines both intuition from skip-gram modeling and non-parametric topic models, which can learn concept representations from both context and topics. Experiments show that our model outperforms other strong baselines in multi-label diagnose classification task, and learns multi-sense representations with better interpretability.

Acknowledgments

Thanks for the valuable instruction and suggestions from Natali and Professor Jian Tang.

This paper is partially supported by the National Natural Science Foundation of China (NSFC Grant Nos. 91646202, 61772039 and 61472006).

References

[1] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1495–1504.

[2] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 41, 2016.

[3] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Medical concept representation learning from electronic health records and its application on heart failure prediction," *arXiv preprint arXiv:1602.03686*, 2016.

[4] J. A. Minarro-Giménez, O. Marín-Alonso, and M. Samwald, "Exploring the application of deep learning techniques on medical text corpora." *Studies in health technology and informatics*, vol. 205, pp. 584–588, 2013.

[5] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza, "Medical semantic similarity with a neural language model," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 1819–1822.

[6] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[7] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.