Clinical Knowledge Graph Embedding Representation Bridging the Gap between Electronic Health Records and Prediction Models

Matthew Wai Heng Chung
Institute of Health Informatics
University College London
London, UK
wai.chung.18@ucl.ac.uk

Jianyu Liu
Institute of Health Informatics
University College London
London, UK
jianyu.liu@ucl.ac.uk

Hegler Tissot
Institute of Health Informatics
University College London
London, UK
h.tissot@ucl.ac.uk

***Abstract*—Learning knowledge embedding representation is an increasingly important technology. However, the choice of hyperparameters is seldom justified and usually relies on exhaustive search. Understanding the effect of hyperparameter combinations on embedding quality is crucial to avoid the inefficient process and enhance practicality of embedding representation along subsequent machine learning applications. This work focuses on translational embedding models for multi-relational categorized data in the clinical domain. We trained and evaluated models with different combinations of hyperparameters on two clinical datasets. We contrasted the results by comparing metric distributions and fitting a random forest regression model. Classifiers were trained to assess embedding representation quality. Finally, clustering was tested as a validation protocol. We observed consistent patterns of hyperparameter preference and identified those that achieved better results respectively. However, results show different patterns regarding link prediction, which is taken as strong evidence that traditional evaluation protocol used for open-domain data does not necessarily lead to the best embedding representation for categorized data.**

***Index Terms*—electronic health records; multi-relational data; knowledge graphs; embedding representation; link prediction; clustering; classification;**

I. Introduction

Domain-specific data sources usually contain high dimensional data which makes data-driven tasks such as knowledge reasoning and inference challenging. In the clinical domain, the storage of patient information in the form of an electronic health record (EHR) allows recording of everything from symptoms to test results and diagnoses, with standardized definitions to ensure consistency. Consequently, the problem of representing multi-relational data has gained more attention in the last decade as long as more knowledge graphs (KGs) become available and useful as supporting resources for a variety of machine learning applications. Knowledge embedding representation (KER) methods are able to learn and operate on the latent feature representation of the constituents and on their semantic relatedness. Thus, KER has the ability to semantically encode multi-relational data in such a way that this latent representation can be efficiently used as a feeding input in subsequent machine learning applications [1].

In open-domain KGs, the heterogeneous nature of the data sources where facts are usually extracted from makes the data typically inaccurate. Moreover, although containing a huge number of triplets, most open-domain KGs are incomplete, covering only a small subset of the true knowledge that they are supposed to represent. In domain-specific KGs, incompleteness results from missing values and cardinality-related inconsistencies that are usually produced by automatic IE processes from unstructured data sources (e.g. clinical notes) [2]. Learning the distributed representation of multi-relational data has been used as an efficient tool to complete and validate knowledge bases without requiring extra knowledge. Knowledge base completion or link prediction (LP) refers to the problem of predicting new links (or new relationships) between entities by automatically recovering missing facts based on the observed ones.

In a knowledge graph $\mathcal{G}$ constructed with a set of facts $\mathcal{S}$ in the form of triples, each triple (h, r, t) has a pair of head h and tail t entities, collectively termed as entities $e \in \mathcal{E}$, connected by a relation $r \in \mathcal{R}$. KG embedding methods in general aim to represent entities and relations as vectors in a continuous k-dimensional vector space $\mathbb{R}$, so that entity embedding vectors $\mathbf{e} \in \mathbb{R}^k$ will be learned together with a relation embedding vector $\mathbf{r} \in \mathbb{R}^k$. By learning how to operate on the latent feature representation of the triple constituents, embedding models are able to use semantic relatedness to enforce the embedding compatibility. Different types of embedding models for KGs have been proposed, with a common goal to improve low-dimensional KG representation as evaluated by specific tasks.

TransE [3] is a baseline translational approach that use simple assumptions to learn vectors $\mathbf{h}, \mathbf{t}, \mathbf{r} \in \mathbb{R}^k$, so that every relation r is a translation between h and t in the embedding space. The pair of embedded entities in a triple (h, r, t) can be approximately connected by $\mathbf{r}$ with low error ($\mathbf{h} + \mathbf{r} \approx \mathbf{t}$). In addition, a plausibility score function for an embedded triple is calculated as the distance between $\mathbf{h} + \mathbf{r}$ and $\mathbf{t}$. HEXTRATO [4] is a knowledge embedding approach design to operate with categorized multi-relational data that extends TransE with ontology-based constraints. In this work, we use

HEXTRATO to assess the quality of embedding representation of clinical datasets by contrasting usual LP evaluation protocol against classification and clustering tasks. We analyze the accuracy effect of hyperparameter choice when evaluating the resulting embedding representation in a classification task, and we conduct non-parametric sensitivity analysis based on feature importance and partial dependence analysis from random forest regression modeling. In addition, we explore clustering as an alternative validation protocol for KER.

We contrast how the combinations of hyperparameters in learning embedding representation of categorized domain-specific clinical knowledge bases reflect on narrowing down the choices for better representation quality in such KGs, so that an exhaustive grid search can be avoided. Complementary evaluation assessment based on Spearman's rank order correlation coefficient is reported. We found some consistent patterns that can inform the choice of hyperparameters. However, the decision should take into consideration the intended application of the embedding representations, as the best combination of hyperparameters to optimize LP tends to differ from that for a subsequent classification task. Although LP and either cluster or classification tasks have low or almost no observable correlation, when directly comparing clustering accuracy against classification F1 scores for specific sets of hyperparameters, we observed consistent patterns in which the correlation is stronger.

II. Method

Although most embedding methods use open-domain multi-relational data for evaluation purposes, such datasets are not usually enriched with extra metadata and are inherently noisy and incomplete. In this work, we aim to utilize clinical multi-relational categorized data constructed with information extracted from EHR systems. Such clinical datasets may not be complete, but the overall task here is not focused on knowledge graph completion. Instead, we aim to learn entity embedding representation that can be used as input for further health-related machine learning applications.

Categorized KGs are created with additional metadata, in which each resulting triple is presented in the form *(c_h:h, r, c_t:t)*, where c_h and c_t represent the types of *h* and *t*, and each relation is restricted by domain and range – e.g., in `(patient:P01, hasGender, gender:male)`, the relation *hasGender* is constrained by the domain *patient* and the range *gender*.

We evaluate a set of strategies aiming to improve KER for categorized data. We present an evaluation protocol for embedding clinical knowledge graphs that comprises the following steps: (a) firstly, we use HEXTRATO to embed categorized entities and relations; (b) secondly, we perform a classification task using the resulting embedding representation; and (c) finally, we evaluate the accuracy of KER by clustering entity embeddings as an alternative validation protocol for the traditional LP metrics. The latter is focused on contrasting the correlation between multiple accuracy results.

TABLE I: Statistics of domain-specific clinical benchmark datasets, given by the number of entities, relations, types, and triples in each dataset split – training (LRN), validation (VLD), tuning (TUN) and test (TST) sets [4].

	EHR Datasets	
# (number of)	**Demographics**	**Pregnancy**
Entities	2,237	3,088
Relations	6	5
Types	7	4
Triples (total)	15,345	20,768
LRN	13,875	14,588
VLD	463	1,997
TUN	475	2,093
TST	532	2,090

A. Datasets

In this work, we focus on clinical datasets[1] obtained from *InfoSaude* (InfoHealth) [5], an EHR system. An overview of each dataset is presented below and statistics are depicted in Tables I and II – in both datasets, all relations have the domain *patient* as head type.

EHR-Demographics comprises a set of 2,185 randomly selected patients who had at least one admission between 2014 and 2016. Each patient is described by a set of basic demographic information, including gender, age (range in years) in the admission, marital status (unknown for about 15% of the patients), education level, and two flags indicating whether the patient is known to be either a smoker or pregnant, and the social groups assigned according to a diverse set of rules mainly based on demographic and historical clinical conditions. Demographic features are represented by *many-to-one* relations, whereas association of each patient to social groups is given by a *many-to-many* relation.

EHR-Pregnancy is a dataset used to identify correlations between pre- and post-clinical conditions on pregnant patients with abnormal pregnancy termination, comprised by a set of 2,879 randomly selected pregnant female patients in which pregnancy was inadvertently and abnormally interrupted before the expected date of birth; each patient is described by age (range in years), known date of last menstrual period (LMP), whether the patient had an abortion (regardless of reason), and a list of ICD-10 (the 10th revision of the International Classification of Diseases) codes [6] registered either before or after the LMP date. This is a dataset mostly comprising *many-to-many* relations that connects patients with corresponding the diagnoses.

B. Clinical KER

We used HEXTRATO [4] to pre-train a set of embedding models for each clinical dataset, which is an embedding approach originally designed for multi-relational categorized data. Unlike other translational models, HEXTRATO is design for domain-specific datasets. It is fundamentally a translational

[1] https://github.com/hextrato/KER/tree/master/datasets/ – we obtained permission from the *InfoSaude* team to use and publish de-identified versions of each dataset.

TABLE II: Relation cardinality in the EHR datasets.

(a) Demographics

Relation	Cardinality	Range	# Triples
hasGender	N:1	gender	2185
ageRange	N:1	interval	2185
hasMaritalStatus	N:1	maritalStatus	1844
hasMaxEducation	N:1	education	1815
isSmoker	N:1	boolean	2185
isPregnant	N:1	boolean	506
isSocialGroup	N:N	socialGroup	4625

(b) Pregnancy

Relation	Cardinality	Range	# Triples
ageYearsWhenLMP	N:1	interval	2879
hadAbortion	N:1	boolean	2879
ageWeeksWhenInterrupted	N:1	interval	2879
ICDBeforeLMP	N:N	ICD	5776
ICDAfterLMP	N:N	ICD	6355

model which extends TransE with four ontology-based constraints that make use of source metadata.

Given a training set S of categorized triplets *(c_h:h,r,c_t:t)*, HEXTRATO learns embedding vectors for entities and relations, so that each categorized entity *c:e* is represented by an embedding vector $e_c \in \mathbb{R}^k$, and each relation *r* is represented by a embedding vector $r \in \mathbb{R}^k$. A score function f_r (Equation 1) represents an L2-norm dissimilarity. We used stochastic gradient descent (SGD) [7] for optimization in order to minimize the margin-based loss function $\mathcal{L}$ (Equation 2) adapted from TransE, where γ is the margin parameter, $\mathcal{S}$ is the set of correct triples, $\mathcal{S}'$ is the set of incorrect triples *(c_h:h,r,c_t:t')*, and $[x]_+ = max(0, x)$.

$$f_r(h_{c_h}, t_{c_t}) = \|h_{c_h} + r - t_{c_t}\|_{l_2} \tag{1}$$

$$\mathcal{L} = \sum_{\substack{(c_h:h,r,c_t:t)\in\mathcal{S} \\ (c_h:h,r,c_t:t')\in\mathcal{S}'}} [\gamma + f_r(h_{c_h}, t_{c_t}) - f_r(h_{c_h}, t'_{c_t})]_+ \tag{2}$$

Pre-processing steps included splitting datasets into 4 subsets for learning (LRN), validation (VLD), tuning (TUN) and testing (TST) as presented in Table I. HEXTRATO uses a *tuning* set to choose the best of multiple replicas independently initialized with random vector representations for each entity and relation. For each model with a distinct set of hyperparameters, 5 replicas were trained and validated based on the LP metrics of predicting t. After going through all training and validation cycles for a maximum of 1000 epochs, the instances are scored on the tuning set for comparison and the best one is selected. Final results are obtained using the test set.

We trained models for all combinations of the following hyperparameters: learning rate $\lambda \in \{0.001, 0.01, 0.1\}$, margin $\gamma \in \{1.0, 2.0, 4.0\}$, embedding dimension $k \in \{8, 16, 32, 64\}$, and the proposed ontology constraints in a cumulative way, including types, isolated values, disjoint groups, and functional relations (correspondingly labelled as *H+T*, *H+TI*, *H+TID*, and *H+TIDF* in Fig. 1). Embedding vectors are initialized with random uniform normalized values [8]. Each training step generates a corresponding corrupted triple for the correct one. Finally, a regularization constraint is imposed on the entity embedding vectors, at the end of each training cycle. This prevents loss minimization by artificially increasing the entity embedding norms.

C. Embedding Evaluation Protocol

KER methods are commonly evaluated on the link prediction task, which involves predicting a correct element that is missing to complete a triple (h, r, t). It usually refers to entity prediction to predict either *h* from the given $(?, r, t)$ or *t* from the given $(h, r, ?)$. Instead of a single outcome, a ranked list of candidate entities is returned. Using tail prediction as an example, in order to derive this list, a similarity score is calculated using the scoring function for every candidate triple (h, r, e) generated from all possible entities e, and ranked in descending order. HEXTRATO follows a similar evaluation protocol for link prediction. For each testing triple, the model predicts *t* given $(c_h : h, r, c_t : ?)$. For each entity in a set of candidates a dissimilarity score is calculated, and the rank of the correct missing entity is recorded.

There are multiple metrics to assess the performance of a model on this task. Mean rank (MR) takes the average of the recorded ranks of all correct entities. Mean reciprocal rank (MRR) calculates this average after taking the reciprocal of the ranks of all correct entities, so it is more robust against outliers than MR, where achieving lower MR or higher MRR are taken as good LP scores. HEXTRATO reports the best MR and competitive MRR scores when compared to other translational models in an open-domain LP task.

D. Classification

In order to formulate a possible embedding quality evaluation task, we propose a set of Multi-layer Perceptron (MLP) classifiers that take selected entity embedding vectors as input to predict a specific label for each entity: (a) within the *EHR-Demographics* dataset, each patient is predicted to be in none or multiple of the sixteen social groups, and (b) within the *EHR-Pregnancy* dataset, each patient is predicted to whether or not having an abortion based on other clinical conditions. Each patient set is randomly assigned to the training set (80%), validation set (10%) or the testing set (10%) while maintaining class proportionality. We do not aim to create the best classifiers, but to contrast the quality of all embedding models, as in how good each one performs when learning representations and resembles the original information such as specific entity classes. Therefore, the classifiers have the same controlled structure and parameters unless necessary variations are necessary for the classification task.

Each classifier is designed with three layers. The first input layer has a number of nodes the same as the input vector dimensions. The hidden layer has two times the nodes of the input layer and a sigmoid activation function. The third output layer has node size according to the number of labels in each classification. For the *Pregnancy*, the task is

TABLE III: Resulting Spearman's correlation coefficients among link prediction (MRR), classification, and clustering tasks.

				Spearman's Coefficients		
EHR Dataset	**Target Type**	**Target Relation**	**# of Classes**	**MRR vs Classification**	**MRR vs Clustering**	**Classification vs Clustering**
Demographics	Patient	hasMaritalStatus	4	-0.124	-0.073	+0.143
		ageStage	6	+0.152	+0.056	+0.327
Pregnancy	Patient	hadAbortion	2	-0.063	-0.178	+0.625
		ageWeeksPregnancyInterrupted	15	-0.495	-0.032	+0.609

a binary classification problem - the output layer has one single node with that outputs a probability distribution. For the *Demographics* dataset, the task is a multi-class classification, so the output layer has sixteen nodes with sigmoid activation function. Each label is independently predicted as true if the output value is above a 0.5 threshold. All the hidden and output layers were trained with a uniform kernel initialization.

A stochastic gradient descent (SGD) optimizer is used for training in all classification tasks with 0.01 learning rate and no decay. The categorical cross-entropy loss is used in the *Pregnancy* dataset, whereas the binary cross-entropy is used in the *Demographics* instead. Datasets have an unbalanced number of instances in each class, which causes the classifiers to predict only for the majority class. To overcome this, we used the class distributions which proportionally add more weights to the instances belonging to the minority classes such that all classes are equally represented in training. In each learning epoch, the model learns from the training set to minimize loss, and then makes prediction on the validation set. Validation accuracy is monitored during training and the model is saved if the metric improved after each epoch. After 5000 epochs, the best model is loaded and evaluated on the test set. Accuracy is given by the weighted-by-class F1 score calculated from the predictions, which takes class imbalance into account.

E. Clustering

We compare whether the clustering accuracy on each target relation is correlated to the accuracy given by the MRR metric from the LP task and accuracy from the classification task. For each dataset, we perform cluster analysis (using K-means) separately on the resulting embedding models, so the number of clustering is pairwise with number of models. The target entity types used for clustering and the set of target labels obtained from relation in each KG are described in Table III.

We compare the clustering result from each target label against the embedding model LP accuracy given by MRR and the accuracy given by the classification task. To calculate the K-means accuracy, we use the predominant target label in each cluster as intra-cluster accuracy (Acc_k). The overall relation clustering accuracy (Acc_r) is taken as the intra-cluster weighted average (Equation 3). The weighted average Acc_r considers the number of entities in order to weight each cluster, so that the degree of influence in which intra-cluster accuracy contributes to the overall accuracy rate depends on the proportion of the target predominant labels T_k in each cluster k regarding the total number of target labels L.

$$Acc_r = \frac{\sum_{k=1}^{4L} Acc_k * \frac{T_k}{L}}{\sum_{k=1}^{N} \frac{T_k}{L}} \tag{3}$$

We use $4 \times L$ target clusters to evaluate L target labels, so that it is expected multiple clusters having the same predominant target label, whereas less representative labels in the KG being kept out of the accuracy means.

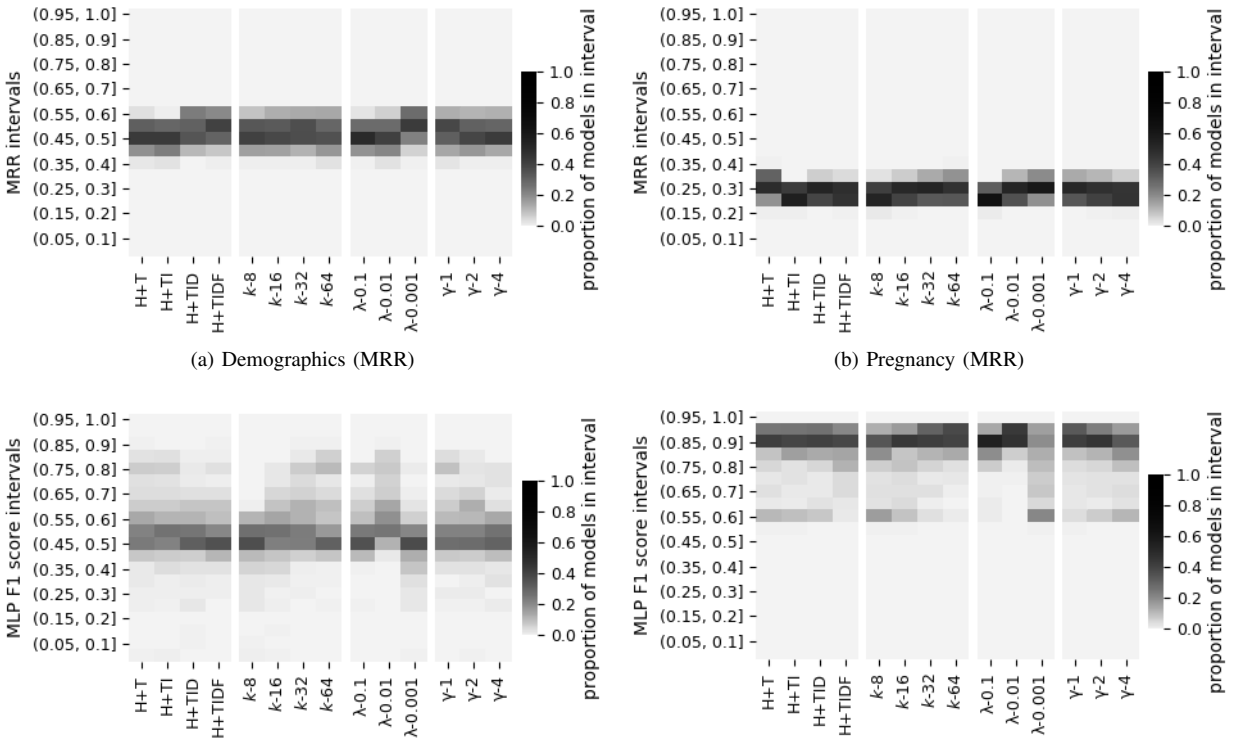
III. Results

We question whether there are associations and patterns between dataset shapes and sizes, hyperparameters and the quality of learned representations. The objective is to understand important factors that determine the effectiveness of representation learning in the clinical domain, so that instead of conducting exhaustive search of the best hyperparameters, we can start with reasonable confined options. We comprehensively trained and tested models of several combinations of hyperparameters to compare and contrast their evaluation results against classification and clustering tasks.

A. Hyperparameter evaluation

The same hyperparameter setup can produce different results in different datasets comparatively to other setups. This implies that at a certain level, dataset feature differentiation can lead to variation in the test score achieved. In addition, we observed none of the particular combinations of hyperparameters significantly outperform the others. Besides specific variety of hyperparameter sets with best performance, many top-ranked models achieved similar scores. We analyze the bigger picture of how these combinations affect the KER learning process and provide some evidence for hyperparameter selection.

In order to identify any patterns between hyperparameters and evaluation metrics in individual KGs, we make use of our comprehensive data to produce heatmaps to visualize the distribution of scores stratified by individual hyperparameters. In Fig. 1, (a) and (b) show the LP score (MRR) distribution heatmaps (separated blocks belong to a distinct hyperparameter, with intervals on the y-axis "closed on the right"), whereas (c) and (d) show the classification (F1) score distribution heatmaps for each corresponding dataset. Each column represents MRR scores across all models when that hyperparameter option is chosen. Comparing between the set of options for a particular hyperparameter shows its overall impact on link prediction. Although less quantitative than conventional sensitivity analyses, these heatmaps provide a unique view of uncertainty in embedding representation



(a) Demographics (MRR)

(b) Pregnancy (MRR)

(c) Demographics (F1)

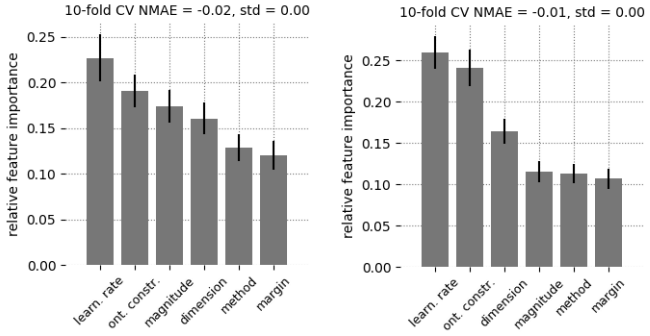
(d) Pregnancy (F1)

Fig. 1: Link prediction (MRR) and classification (F1) score distribution heatmaps by hyperparameter for each dataset.

learning with respect to hyperparameter choices in all possible combinations.

There are three important pieces of information we can acquire from the heatmaps. Firstly, they show the highest achievable MRR interval and the proportion of models in it. For example, in *Demographics*, *H+TIDF* set of ontology constraints has many more models achieving 0.55 to 0.60 MRR than the other two sets, which is the same observed for learning rate equals 0.001. Secondly, we can look at the spread of the distributions, where the heatmap illustrates model concentrations and their relative positions in the distributions. For example, different margins γ in *Demographics* and *Pregnancy* are similar in MRR spread, though γ = 1.0 gives better concentration. Finally, some consistent patterns can be observed across the heatmaps: (a) for ontology constraints, *H+TI* rarely reach the highest MRR interval; (b) higher dimensions usually allow models to reach higher MRR although 64 is not necessarily better than 32; (c) for learning rate, 0.1 is inferior in most of the scenarios; and (d) a higher margin (e.g. $\gamma > 4.0$) is generally less preferable. The latter two suggest that, instead of helping within the learning process, higher values of such hyperparameters produce too much noise pushing the entities beyond the surface of the hyperspace, inducing constant regularization, which has a negative effect on the accuracy.

Lastly, in order to understand the contribution of each hyperparameter at the broader context of all trained models,



(a) Demographics

(b) Pregnancy

Fig. 2: Random forest regression models for EHR datasets predicting MRR, with feature importance for all dimensions. Error bars indicate the standard deviation. "Ont. constr." stands for ontology constraint.

we performed a sensitivity analysis with a random forest regression model [9]. Our findings suggest that, before stratifying the data into different dimensions, learning rate is the most important hyperparameter for MRR prediction, remaining highly important in both datasets (Fig. 2). The importance of learning rate can be noticeably attributed to the negative effect of high value for $\lambda = 0.1$.

B. Clustering and Classification tasks

For each dataset, we used the resulting embeddings and labels for each entity as the input for the clustering model in order to obtain the clustering result accuracy of each embedding model. Then, we compared the accuracy from both evaluation protocols by correlating MRR accuracy against clustering accuracy. We report Spearman's rank order correlation coefficient between each method in Table III.

By comparing the MRR and clustering accuracy, we found correlations are consistently weak for both datasets. Unexpectedly, for *Demographics* and *Pregnancy* datasets, evaluation metrics are mostly uncorrelated to each other, and justification is a two-fold: (a) in both datasets we observed high accuracy repeatedly resulting from clustering, which is not consistent with relatively low accuracy given the LP metrics; alternatively (b) both datasets are predominantly compound by *many-to-many* relations, whereas we used *many-to-one* relations to perform the proposed clustering protocol, and the predominant relations can be overlaying the semantic relationship contribution given by the other relations. Finally, correlations between MRR and clustering are expected to be positive. However, we found contradictory results when analyzing more specific sets of hyperparameters in both EHR datasets, which led us to further reflect on whether certain shapes of datasets suit better for distinct evaluation protocols.

Spearman's coefficient was also used to assess how the classifier scores correlate with the LP metrics. Notably once again, LP metrics have very weak or no observable correlation with *Demographics* and *Pregnancy* classifier accuracy. The Spearman rank-order correlation coefficients between MRR and F1 score are -0.19 for *Demographics* and +0.21 for *Pregnancy*. This may be caused by similar justification as the ones given for clustering. Further analysis in each dataset is required in order to identify whether their high proportion of triples with *many-to-many* relations or the triple count of each dataset may also be a strong determinant.

A summary of hyperparameters choices that have the highest average predicted F1 score along the classification task includes: (a) the ontology constraint set of *H+TID* topped in many partial dependence analyses, while *H+TIDF* is much less preferred; (b) 32 embedding dimensions is the best or second-best model and doubling k to 64 only has marginal benefits; (c) regarding learning rate, 0.1 is likewise undesired, however, the overall preference for 0.01 is the opposite of presented for the MRR random forest model; and (d) the learning margin $\gamma = 1.0$ is predominantly preferable, whereas 2.0 can occasionally outperform at higher dimensions.

IV. Conclusions

KER models with the best link prediction accuracy can result from different combinations of hyperparameters and rely on training dataset shapes or sizes. The primary aim of this work is to evaluate whether link prediction accurately reflects the quality of the resulting embeddings for categorized data, and whether they can be efficiently used in subsequent machine learning tasks instead of KG completion.

We analyzed the MRR distribution of all models that share a particular hyperparameter variant in heatmaps. Additionally, we conducted a non-parametric sensitivity analysis by random forest regression modeling that takes into consideration models of all combinations. It has the strength of modeling non-linear relationship and that the feature importance can be easily calculated. This can reflect the relative influence of changing individual hyperparameter options on predicted MRR. However, this approach assumes hyperparameter independence which may be violated.

Experimental results show that LP metrics might not always reflect quality of the embeddings from categorized data intended for subsequent machine learning tasks. Alternative metric based on the K-Means clustering was tested. In general, clustering accuracy has low or almost no observable correlation with LP metrics. However, although the observed correlations between LP and either clustering or classification tasks are weak, this correlation tends to be stronger when directly comparing clustering accuracy and classification F1 scores for specific sets of hyperparameters. We observed some consistent patterns that can inform the choice of hyperparameters. The decision should take into consideration the intended application of the embedding representations. Most importantly, the best combination of hyperparameters to optimize LP performance tends to differ from that for a subsequent classification task.

References

[1] Z. Liu, M. Sun, Y. Lin, and R. Xie, "Knowledge representation learning: A review," *Journal of Computer Research and Development*, vol. 53, no. 2, p. 247, 2016. [Online]. Available: http://crad.ict.ac.cn/EN/abstract/abstract3099.shtml

[2] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 10, pp. 1411–1428, Oct. 2006. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2006.152

[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2787–2795.

[4] H. Tissot, "HEXTRATO: Using ontology-based constraints to improve accuracy on learning domain-specific entity and relationship embedding representation for knowledge resolution," in *IC3K 2018 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, vol. 1. SciTePress, 2018, pp. 72–81.

[5] H. Tissot and R. Dobson, "Identifying misspelt names of drugs in medical records written in portuguese," *HealTAC-2018: Unlocking Evidence Contained in Healthcare Free-text*, 2018.

[6] WHO, *ICD-10: International Statistical Classification of Diseases and Related Health Problems / World Health Organization*, 10th ed. World Health Organization Geneva, 2004.

[7] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 09 1951.

[8] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, Y. W. Teh and D. M. Titterington, Eds., vol. 9, 2010, pp. 249–256.

[9] F. Hutter, H. Hoos, and K. Leyton-Brown, "An efficient approach for assessing hyperparameter importance," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32-1. Bejing, China: PMLR, 6 2014, pp. 754–762. [Online]. Available: http://proceedings.mlr.press/v32/hutter14.html