Medical Schema Matching using Knowledge Graph Embedding

Chaoyu Gao
Southeast University, China
cygao@seu.edu.cn

Tianxing Wu*
Southeast University, China
tianxingwu@seu.edu.cn

Shengqi Jing
The First Affiliated Hospital of Nanjing Medical University, China
jingshenqi@jsph.org.cn

Yuxiang Wang
Hangzhou Dianzi University, China
lsswyx@hdu.edu.cn

ABSTRACT

Heterogeneous medical schema matching is important to realize the data sharing between medical databases. Traditional medical schema matching algorithms which usually employ hand-crafted features created by medical experts fail to handle increasingly giant data with the continuous development of medical digitization. In this paper, we propose a new medical schema matching method which extracts equivalent relations between columns of relational tables in heterogeneous databases using knowledge graph embedding. Our preliminary experiments on a real-world medical dataset show the superiority and competitiveness of our proposed method.

KEYWORDS

Medical databases, Medical Knowledge, Medical Schema Matching, Knowledge Graph Embedding

ACM Reference Format:
Chaoyu Gao, Tianxing Wu, Shengqi Jing, and Yuxiang Wang. 2022. Medical Schema Matching using Knowledge Graph Embedding. In *Proceedings of EEKE '22: The 3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE '22).* ACM, New York, NY, USA, 5 pages. https://doi.org/TBD

1 INTRODUCTION

In recent years, the continuous development of medical informatization makes a large number of diagnosis and treatment processes recorded in databases, which contain much valuable medical knowledge. However, different medical institutions have different schemas in their management system. Since the data stored in medical databases are of various types and forms, it is difficult to directly fuse data which seriously affects the analysis and application of medical big data. Thus, it becomes an urgent problem to effectively fuse multi-source data from different medical databases.

To fuse heterogeneous databases, lots of efforts have been made in schema matching [1], which is the process of capturing correspondence between columns of different relational tables. Some researches [2, 3] leverage the experience of experts and statistical machine learning methods to design hand-crafted features and create regular expressions for schema matching, which requires sufficient labeled data, and the corresponding annotation costs. With the great success of pre-trained language models, one line of researches [1, 4] attempts to vectorize instances by pre-trained language models, and make semantic comparison with vector similarities to get the equivalent relations between columns.

Although the previous researches have achieved good results, they have the following problems when directly applied in the medical field. Firstly, relations among medical instances are complicated, which makes it difficult for matching medical schemas using pre-trained language models. Secondly, medical schema matching lacks domain knowledge, which lowers the quality of matching results.

To address the above problems, we propose a new medical schema matching method using knowledge graph embedding which incorporates medical domain knowledge. Firstly, the direct mapping method is used to transform structured databases into corresponding knowledge graphs, which better characterizes the contextual information of entities and properties. Secondly, we take the text descriptions of entities from the Chinese medical encyclopedia[1](an online medical encyclopedia describing medical entities) as background knowledge, and utilize the pre-trained language model BERT [5] to train vector representations of entities. Finally, knowledge graph embedding is used to discover the equivalent relations between properties across heterogeneous knowledge graphs, i.e., the equivalent relations between columns of original relational tables.

Our contributions are summarized below:

- We first propose a medical **S**chema **M**atching method by **K**nowledge **G**raph **E**mbedding (KGESM) to capture equivalent relations between columns of relational tables in heterogeneous databases;
- We propose a new medical entity description embedding based on BERT by encoding medical entities and their description;
- We conduct preliminary evaluations on a real-world dataset in terms of precision, recall, and F1-score, and the results

*Corresponding Author.

EEKE '22, June 20-–24, 2022, Online

ACM ISBN TBD…$TBD
https://doi.org/TBD

[1] http://www.a-hospital.com/

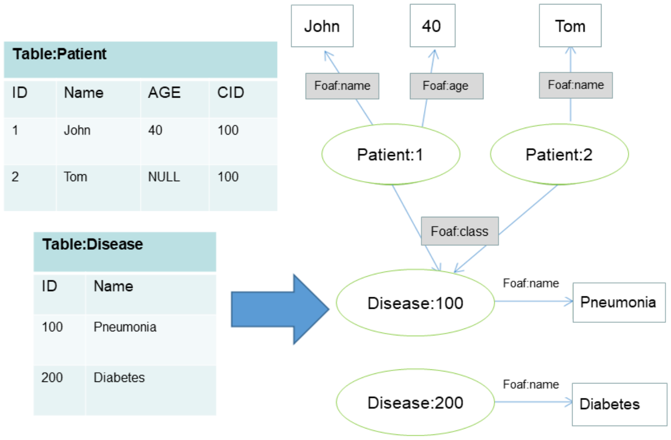

Table:Patient

ID	Name	AGE	CID
1	John	40	100
2	Tom	NULL	100

Table:Disease

ID	Name
100	Pneumonia
200	Diabetes

Figure 1: The example of direct mapping.

show the superiority and competitiveness of our proposed method.

2 THE PROPOSED APPROACH

We begin our modeling with transformation from structured databses to knowledge graphs.

2.1 Direct Mapping

Considering the complexity of concepts and properties in the medical field, the schema matching methods on structured databases are difficult to handle medical data, while the knowledge graph could more comprehensively characterize the contextual information of concepts and properties in the form of graphs. Therefore, the schema matching methods based on knowledge graph could improve the knowledge fusion of concepts and properties in medical field. Referring to the Direct Mapping standard established by W3C[2], we propose a rule-based method to transform medical structured databases into medical knowledge graphs for knowledge fusion. The details are as follows:

- Table names of the structured databases are used as concepts;
- Columns of tables are regarded as properties;
- Rows of tables are regarded as instances;
- Cell values in each row are literals or instances (if the column is a foreign key).

As shown in figure 1, the structured databases could be transformed into a medical knowledge graph composed of triples.

2.2 Medical Entity Description Embedding

The existing text-based pre-trained language model (PTLM) lacks medical knowledge to capture the semantic information of low-frequency medical professional words, while the Chinese medical encyclopedia contains rich structured information, which could greatly improve the learning performance of the existing PTLM. To better represent concepts and properties in the medical field, we propose a medical entity description embedding framework (MEDE) based on the pre-trained language model. As shown in figure 2, MEDE consists of two steps: firstly, the medical entity is used as a query to extract a related description from the Chinese medical encyclopedia, then the entity and description are concatenated as the input. Secondly, the input will be encoded by BERT encoder, and then pooled into the vector as the output of the encoder.

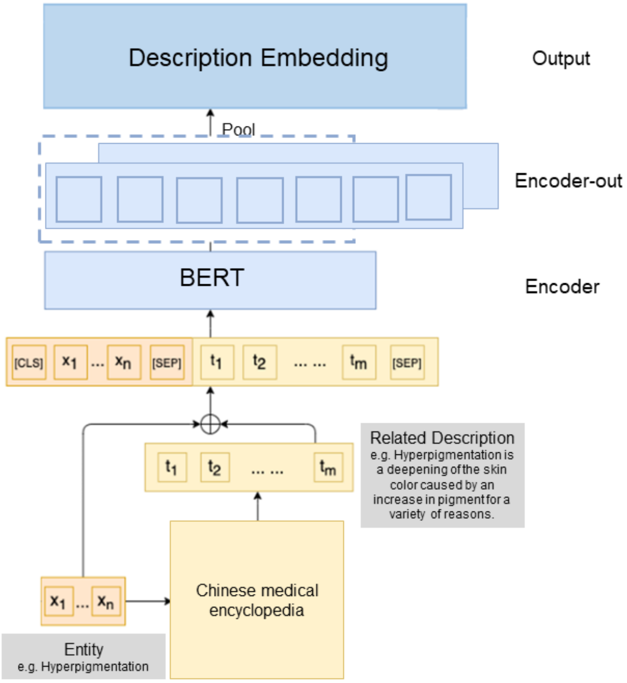

Figure 2: The medical entity description embedding framework. x_i in $(x_1, ..., x_n)$ and t_j in $(t_1, ..., t_m)$ respectively denote each word in the medical entity and its description.

Input. The input to MEDE consists of the medical entity and its relevant description from the Chinese medical encyclopedia. In this paper, we follow the standard method where the tokens are surrounded respectively by [CLS] and [SEP] on the left and right. For example, if the input entity and description are "*Hyperpigmentation*" and "*Hyperpigmentation is a deepening of the skin color caused by an increase in pigment for a variety of reasons.*", the input to our encoder would be: [CLS]+ "*Hyperpigmentation*" +[SEP]+ "*Hyperpigmentation is a deepening of the skin color caused by an increase in pigment for a variety of reasons.*" + [SEP].

Text encoder (BERT) details. For the MEDE model, we use the pretrained model BERT which is proposed by Devlin et al. [5]. BERT is a multi-layer bidirectional Transformer encoder, and we use its original implementation, the BERT-base model. Considering that the deeper Transformer layer may be more related to the pretraining tasks mask language model and next sentence prediction, we remove the output of the last Transformer layer and only take the outputs of the second and third layers. For the outputs of these two layers, we first use the pooling layer to further integrate the information from different channels with two results connected

[2] https://www.w3.org/TR/rdb-direct-mapping/

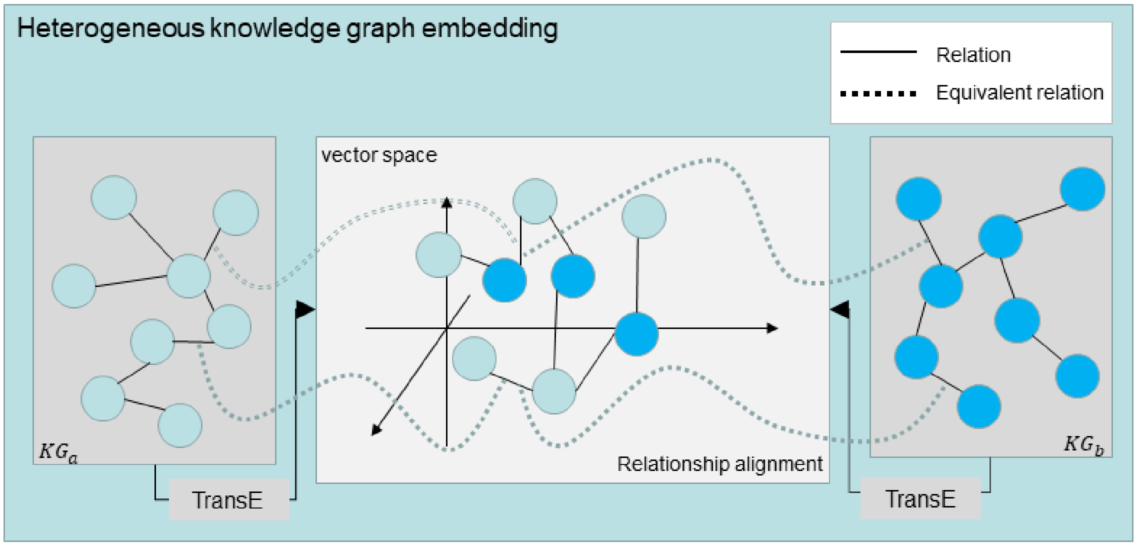

Figure 3: The schema matching framework. KG_a **and** KG_b **denote different knowledge graphs which are embedded by TransE.**

together, and then adopt the average pooling strategy to output the vector (768×1) as the output of the encoder.

2.3 Schema Matching

In this subsection, we present our schema matching framework using knowledge graph embedding. As shown in figure 3, the proposed framework consists of two modules: heterogeneous knowledge graph embedding module, and relation alignment module. Specially, heterogeneous knowledge graph embedding module learn knowledge graph embeddings; relation alignment module captures the equivalent relations from different knowledge graphs, and discovers new equivalent relations through iterative training.

Heterogeneous knowledge graph embedding. Assuming E_a and R_a are respectively employed to represent entity and relation sets of knowledge graph G_a; E_b and R_b are respectively employed to represent entity and relation sets of knowledge graph G_b. We firstly initialize entities and relations of heterogeneous knowledge graphs using vectors obtained in Section 2.2, then employ the basic translation-based method TransE [6] for the involved knowledge graphs, which embeds different knowledge graph in different spaces, and calculates their common loss function. The translation score could be obtained as below:

$$f_r(h,r) = \sum_{G \in \{G_a, G_b\}_{(h,r,t) \in G}} ||h + r - t|| \tag{1}$$

where $T = (h,r,t)$ denotes a triple in G such that $h, t \in E$ and $r \in R$. The loss function is given as below:

$$S_e = \sum ||f_r(h,r) - f_r(\tilde{h},\tilde{r}) + \gamma|| \tag{2}$$

where $(\tilde{h}, r, \tilde{t})$ is the negative example obtained by randomly replacing the header entity or tail entity in (h,r,t) with another entity, and λ is the parameter describing the boundary between positive and negative examples.

Relation alignment. The objective of relation alignment is to capture the transformation between the vector spaces. We adopt the linear transformation based technology, which has pretty good effectiveness among the alignment methods based on Knowledge Graph Embedding [7]. The loss function is given as below:

$$S_t = \sum_{r,\bar{r} \in IR(R_a,R_b)} ||M_{ij}r - \bar{r}|| \tag{3}$$

where $(r, \bar{r})$ denotes the equivalent relation pair in the equivalent relation set $IR(R_a, R_b)$, and the 768×768 square matrix M_{ij} could be employed as a linear transformation on relation vectors from R_a to R_b, given 768 as the dimensionality of vectors obtained in Section 2.2. To combine the above two loss function, we minimize the following loss function: $S_{kg} = S_e + \alpha S_t$, where α is a hyperparameter that weights S_e and S_t.

As for training details, we initialize matrices using random orthogonal initialization [8], then optimize the loss function using stochastic gradient descent [9]. At each iteration we employ the K-Nearest Neighbor algorithm [10] to discover new equivalent relations, which would be obtained only if the distance within a certain threshold.

3 EXPERIMENT

3.1 Datasets

We evaluated our proposed method on a real-world dataset, which consists of the medical data information from Jiangsu Province Hospital relevant medical treatment combination (Hospital J and Hospital K denote them for privacy) in the third quarter of 2020, manually labeled equivalent relation pairs across the heterogeneous medical data, and medical entity descriptions of the Chinese medical encyclopedia. Data of Hospital J and Hospital K were respectively transformed into medical knowledge graph J and K through direct mapping. Table 1 shows the statistics of the knowledge graphs. We invite medical experts to obtain high-quality labeled data manually

Table 1: The statistics of the knowledge graphs

		Num
Knowledge graph J	Entity	3,625
	Triple	66,845
	Property	436
Knowledge graph K	Entity	2,830
	Triple	56,517
	Property	335

from tabular data, which is composed of 287 pairs of equivalent relations. We split the data into 200 for training to fine-tune the model and 87 for testing randomly.

3.2 Experiment Settings

We evaluated the effectiveness of the proposed model in the testing set. We compared our method with the following baselines:

- **SMGR [3]**: SMGR adopts the strategies that combine the strengths of Google similarity as a web semantic and regular expression as pattern recognition to identify 1-to-1 schema matching.
- **ISMW [4]**: ISMW proposed an instance-based schema matching approach using the Word2Vec as the semantic similarity to capture accurate relations between heterogeneous data.
- **REMA [1]**: Graph Embeddings based Relational Schema Matching obtains semantically relevant columns based on context information extracted from the data table to generate pattern matches.

For the training details, we used the BERT implementation of transformers library[3], and employed the Chinese version for preprocessing the BERT-base model. We perform hyperparameter tuning for our proposed model in terms of learning rate among $\{3e-5, 2e-5\}$, batch size among $\{32, 64\}$. The fine-tuning for the BERT-base model took 10 epochs. The proposed model used in our experiments was trained with a batch size of 64, the Adam as optimizer with a starting learning rate of 2e-5 (linearly decayed throughout training), and the dropout rate of 0.1. As for the baseline models, we used the default hyperparameters.

3.3 Experimental Results

Our proposed model, KGESM achieves the best performance in terms of precision, recall, and F1-score, which is shown in Table 2. Although SMGR uses Google similarity and regular expressions as the basis of schema matching, the traditional statistical machine learning method SMGR needs large-scale labeled data, which is undoubtedly missing in the medical field, and causes that SMGR being inferior to other methods in both accuracy and recall. In addition, ISMW performs marginally better and verifies the excellent performance of the pre-trained language models. However, owing to the influence of the unknown words which is common in the medical field, it is hard for Word2Vec to encode medical concepts and properties. While BERT uses word-based coding, which avoids the influence of unknown words, so the KGESM model performs better than ISMW model. As for REMA, it also performs well, and shows the power of exploiting contextual information through graph embedding. However, because of large presence of contractions and misspellings as well as vocabulary mismatch between medical databases, the precision of REMA is affected. Instead, KGESM employs medical knowledge from the Chinese medical encyclopedia to correct these errors in the medical databases and effectively improve the results of schema matching.

Table 2: The evaluation results of all comparison models (%).

model	Precision	Recall	F1-Score
SMGR	79.8	87.2	83.3
ISMW	81.5	87.8	84.5
REMA	83.4	90.6	86.9
KGESM	**85.9**	**95.2**	**90.3**

4 CONCLUSION AND FUTURE WORK

In this paper, we combine the medical domain-specific knowledge and the knowledge graph embedding model to handle the task of medical schema matching. We first propose a medical schema matching method (KGESM) by knowledge graph embedding to capture equivalent relations between columns of relational tables in heterogeneous databases. With the direct mapping method, we extract knowledge from relational databases to construct medical heterogeneous knowledge graphs, then we leverage the Chinese medical encyclopedia to learn better representation of medical concepts and properties. Finally, our proposed schema matching framework employs the knowledge graph embedding to discover the equivalent relations. Experimental results show the high quality of KGESM on the real-world dataset.

As for the future work, we will explore to mine complex equivalent relations (e.g., 1-to-n and m-to-n matches) between properties in medical schema matching. In addition, we plan to develop the framework for automatically matching multiple medical schemas simultaneously.

REFERENCES

[1] Christos Koutras, Marios Fragkoulis, Asterios Katsifodimos, and Christoph Lofi. Rema: Graph embeddings-based relational schema matching. In *EDBT/ICDT Workshops*, 2020.

[2] Benjamin Zapilko, Matthäus Zloch, and Johann Schaible. Utilizing regular expressions for instance-based schema matching. In *OM*, 2012.

[3] Osama Mehdi, Hamidah Ibrahim, and Lilly Affendey. An approach for instance based schema matching with google similarity and regular expression. *International Arab Journal of Information Technology (IAJIT)*, 14(5), 2017.

[4] Kenji Nozaki, Teruhisa Hochin, and Hiroki Nomiya. Semantic schema matching for string attribute with word vectors and its evaluation. *Int. J. Networked Distributed Comput.*, 7(3):100–106, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[7] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

[8] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint*

[3] https://huggingface.co/transformers/

arXiv:1312.6120, 2013.

[9] D Randall Wilson and Tony R Martinez. The general inefficiency of batch training for gradient descent learning. *Neural networks*, 16(10):1429–1451, 2003.

[10] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 986–996. Springer, 2003.