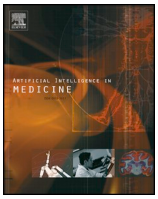
Artificial Intelligence In Medicine 103 (2020) 101817

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Real-world data medical knowledge graph: construction and applications

Linfeng Li[a,b,1], Peng Wang[c,d,1], Jun Yan[b], Yao Wang[b], Simin Li[b], Jinpeng Jiang[b], Zhe Sun[b], Buzhou Tang[e], Tsung-Hui Chang[f], Shenghui Wang[a], Yuting Liu[g,*]

[a] *Institute of Information Science, Beijing Jiaotong University, Beijing, China*
[b] *Yidu Cloud Technology Inc., Beijing, China*
[c] *College of Computer Science, Chongqing University, Chongqing, China*
[d] *Southwest Hospital, Chongqing, China*
[e] *Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China*
[f] *The School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China*
[g] *School of Science, Beijing Jiaotong University, Beijing, China*

ARTICLE INFO

Keyword:
real-world data
medical knowledge graph
CDSS
quadruplet
PSR

ABSTRACT

Objective: Medical knowledge graph (KG) is attracting attention from both academic and healthcare industry due to its power in intelligent healthcare applications. In this paper, we introduce a systematic approach to build medical KG from electronic medical records (EMRs) with evaluation by both technical experiments and end to end application examples.

Materials and Methods: The original data set contains 16,217,270 de-identified clinical visit data of 3,767,198 patients. The KG construction procedure includes 8 steps, which are data preparation, entity recognition, entity normalization, relation extraction, property calculation, graph cleaning, related-entity ranking, and graph embedding respectively. We propose a novel quadruplet structure to represent medical knowledge instead of the classical triplet in KG. A novel related-entity ranking function considering probability, specificity and reliability (PSR) is proposed. Besides, probabilistic translation on hyperplanes (PrTransH) algorithm is used to learn graph embedding for the generated KG.

Results: A medical KG with 9 entity types including disease, symptom, etc. was established, which contains 22,508 entities and 579,094 quadruplets. Compared with term frequency - inverse document frequency (TF/IDF) method, the normalized discounted cumulative gain (NDCG@10) increased from 0.799 to 0.906 with the proposed ranking function. The embedding representation for all entities and relations were learned, which are proven to be effective using disease clustering.

Conclusion: The established systematic procedure can efficiently construct a high-quality medical KG from large-scale EMRs. The proposed ranking function PSR achieves the best performance under all relations, and the disease clustering result validates the efficacy of the learned embedding vector as entity's semantic representation. Moreover, the obtained KG finds many successful applications due to its statistics-based quadruplet.

where $N_{co}^{\min}$ is a minimum co-occurrence number and R is the basic reliability value. The reliability value can measure how reliable is the relationship between S_i and O_{ij}. The reason for the definition is the higher value of $N_{\mathrm{co}}(S_{i,} O_{ij})$, the relationship is more reliable. However, the reliability values of the two relationships should not have a big difference if both of their co-occurrence numbers are very big. In our study, we finally set $N_{co}^{\min} = 10$ and $R = 1$ after some experiments. For instance, if co-occurrence numbers of three relationships are 1, 100 and 10000, their reliability values are 1, 2.96 and 5 respectively.

1. Introduction

Knowledge graph (KG) has received a lot of attention in recent years. In 2012, Google applied the KG in search engines; since then, the knowledge graph has been used in many application fields [1]. In the medical domain, the knowledge graph is the fundamental component

[*] Corresponding author at: Beijing Jiaotong University, Beijing, 100044, China.
E-mail address: ytliu@bjtu.edu.cn (Y. Liu).

[1] Linfeng Li and Peng Wang contributed equally.

https://doi.org/10.1016/j.artmed.2020.101817
Received 23 September 2019; Received in revised form 19 December 2019; Accepted 4 February 2020

for artificial intelligence (AI) aided medical systems, such as clinical decision support systems (CDSSs) for diagnosis and treatment [2–5], self-diagnosis utilities to assist patient evaluating health condition based on symptoms [6,7].

The KG is a graph-based knowledge representation and organization method, which uses a set of *subject-predicate-object* triplets to represent the various entities and their relationships in a domain. Each triplet is called as a *fact* as well. In KG, nodes represent entities and edges represents relationships between entities. For example, '*Parkinson's Disease*' and '*Tremor*' are concrete entities of type *Disease* and *Symptom* in the medical domain. Given that '*disease_related_symptom*' is a relationship between disease and symptom entities, '*Parkinson's Disease disease_related_symptom Tremor*' is a triplet to represent that '*Tremor*' is a related symptom with '*Parkinson's Disease*'.

Most previous works tried to construct the KG from medical articles, some of them are constructing manually and others are automatically. However, manually constructing KG requires tremendous clinical expert time and effort. For example, it was reported that about fifteen person-years are required to build the Internist-1/QMR knowledge base [8]. Automatically constructing KG from articles is a challenging work as the materials are almost unstructured, which is difficult to understand by computer.

In recent years, thanks to the rapid progress of big data and natural language processing (NLP) technologies, automatically mining knowledge from electronic medical records (EMRs) becomes a promising research trend [9–16,20,21],. Learning medical KG from EMRs is less labor-consuming and more feasible than learning from articles. More importantly, the statistical properties of real-world data based KG make it easier to use.

A lot of papers introduces EMR processing algorithms, including named entity recognition (NER) [9–13], entity normalization [14–16], relation extraction/ranking [26] and graph embedding [20,21]. However, there still lacks an efficient and systematic procedure to build medical KG from EMR data end to end. This paper aims to establish a systematic procedure to construct the medical KG from large-scale EMRs. The study is performed on a big-data platform of a 3A-class hospital in China and the constructed KG contains 9 entity types, totally 22,508 entities. Based on the data, we propose a new quadruplet structure to represent a KG fact, in contrast to the classical triplet structure, and build a total of 579,094 quadruplets. PrTransH [21] is used to train embedding vector for each entity and relation. To evaluate the effectiveness, the obtained KG is applied to several practical applications including CDSS, information retrieval and knowledge transfer with neural networks. At the end of this paper, conclusion of this paper and the prospect of the future work are drawn.

2. RELATED WORK

2.1. Entity recognition and normalization

Luo L et al. [11] proposes an attention-based bidirectional long short-term memory with a conditional random field layer (Att-BiLSTM-CRF), to document level chemical NER. The approach leverages document-level global information obtained by attention mechanism to enforce tagging consistency across multiple instances of the same token in a document. Zhang Y et al. [12] implements the BiLSTM-CRF model to simultaneously recognize five types of clinical entities on Chinese EHR corpus. Ji B et al. [13] proposes a hybrid approach that is composed of BiLSTM-CRF model, drug dictionary and post-processing rules to perform medical NER on Chinese EMRs. The experiment result shows that the approach achieves state-of-the-art performance.

Li H et al. [14] proposes a CNN-based ranking approach for biomedical entity normalization. Lou Y et al. [15] proposes a transition-based model for joint disease entity recognition and normalization, the result shows that the joint model improves the performance of disease entity recognition and normalization significantly. Fei L et al. [16] investigates the effectiveness of BERT-based models for the entity normalization task in the biomedical and clinical domain, which proves that BERT-based normalization models outperformed some state-of-the-art systems.

2.2. Related-entity ranking

We find that there are very limited works about related entity ranking in KG, this may because KG is a relatively new direction, the related works still focuses on mining entities and relations for KG. In text information retrieval domain, Salton G et al. [17] proposes the well-known Term Frequency Inverse Document Frequency (TF/IDF) algorithm in automatic text retrieval, till today it is still the fundamental function to rank documents for a given query. Vechtomova O et al. [18] compares several unsupervised rank functions and TF/IDF performed best between the single algorithms. Kang C et al. [19] proposes a supervised learning model for entity ranking in the Web search domain.

2.3. Graph embedding

Graph embedding aims to map the symbolic entities and relations of one graph into a continuous low-dimension vector space. Graph embedding is an effective semantic representation method for entities and relations. Several translation-based algorithms are proposed in the KG of the general domain. Bordes A et al. [22] proposes TransE which is a translation-based algorithm to model the triplets in KGs. Its target is to make $\boldsymbol{s} + \boldsymbol{p} \to \boldsymbol{o}$, where $\boldsymbol{s}$, $\boldsymbol{p}$ *and* $\boldsymbol{o}$ indicates the embedding vector of subject, predicate and object, respectively. In Wang Z et al. [23], the subject and object embedding vectors are mapped to the relation dependent hyperplane, making it possible to project one entity into different projection vectors in different relations. Lin Y et al. [24] introduces a relation-specific space instead of a hyperplane. Li L et al. [21] proposes an improved algorithm to learn representation vectors from probabilistic medical KG and applied the embedding on link prediction. Wang M et al. [20] trains graph embedding on a patient-disease-medicine KG and applied the embedding on medicine recommendation.

2.4. Medical knowledge graph construction and its applications

There are many related works about building medical KG directly from EMR in recent years. Finlayson SG et al. [25] builds a graph of drugs, diseases, procedures and devices, totally 1 million clinical concepts from 20 million clinical notes, the co-occurrence matrix is also provided. Rotmensch M et al. [26] builds a graph of 156 diseases and 491 symptoms based on 273,174 patient visits to the emergency department. Zhao C et al. [27] constructs an EMR-based medical knowledge network (EMKN) by extracting the medical entities, which contains 6733 nodes and 154,462 edges. Based on the network, a diagnosis model based on only symptoms is proposed. Zhao C et al. [28] develops a new EMR-driven medical knowledge representation and inference system with EMKN, Markov random field (MRF) and representation learning techniques. Furthermore, the system can use the current condition of a patient as input to obtain corresponding recommendations for medical tests, possible diseases, and treatment plans. Shen Y et al. [29] constructs a clinical Bayesian network directly from 10,000 deidentified EMRs, the network contains three entity types, i.e. disease, symptom, risk factor.

Compared with the previous works on building medical KG from EMR, our work have several obvious novelties: 1) covering the 9 entities often used in medical AI systems than the few entity types in previous works; 2) the input EMR is large scale and covers all departments of a 3-A hospital than small amount of data in previous works; 3) we introduce the systematic procedure to build medical KG from EMR data than only focusing on specific steps in previous works.

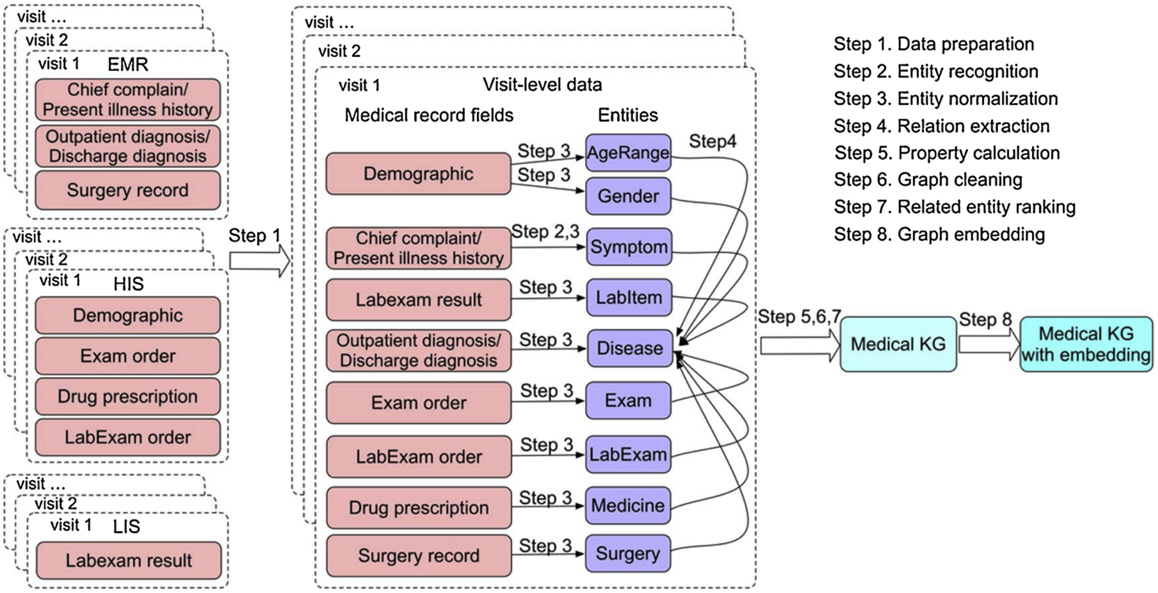


Fig. 1. Proposed systematic procedure of medical KG construction.

3. METHOD

In this section, we develop a systematic procedure to build the medical KG from large-scale EMRs. The procedure, as shown in Fig. 1, involves 8 main steps, which are 1) data preparation, 2) entity recognition, 3) entity normalization, 4) relation extraction, 5) property calculation, 6) graph cleaning, 7) related-entity ranking, and 8) graph embedding, respectively. Here we emphasize that the steps 4), 5), 7) and 8) usually require much practical experience on large-scale EMRs and thus are rarely mentioned in the literature.

3.1. Data Preparation

In this paper, the medical KG is built based on the medical data from 2015 to 2018, which includes 16,217,270 de-identified visits of 3,767,198 patients in total. The data are collected by the data process and application platform (DPAP) at the Southwest Hospital in China, which is deployed in a private cloud based on distributed computing architecture. It aggregates medical data from EMR system, hospital information system (HIS), laboratory information system (LIS) and radiology information system (RIS), etc.

The medical data are organized at the level of visit. Specifically, the visit-level data are collected from all subsystems for each individual visit, including chief complaint, present illness history from EMR, examination orders, drug prescriptions from HIS, laboratory exam results from LIS and so on.

3.2. Entity Recognition

In the medical domain, the named entity recognition (NER) aims to identify medical entities from the free texts, such as diseases, medicines, and symptoms, etc. In our study, symptoms need to be extracted from chief complaint and present illness history by the NER methods, all the other entities can be directly extracted from the structured field of EMRs (as illustrated in Fig. 1).

One of the state-of-the-art methods for NER is BiLSTM-CRF [11,12], which, however, suffers from two major drawbacks, namely, 1) we can't force the deep neural network to recognize an unrecognized entity, even confirmed by a physician; 2) some symptoms cannot be well recognized due to vague symptom boundary. For example, 'chest pain and discomfort (胸痛不适)' indicates two symptoms, i.e., 'chest pain (胸痛)' and 'chest discomfort (胸部不适)'. Therefore, a hybrid model is implemented to resolve these issues. Specifically, it is composed of three components, namely, vocabulary-based bidirectional maximum matching (BMM), BiLSTM-CRF model and pattern recognizer. The BMM is used to recognize the known entities defined in the vocabulary; the BiLSTM-CRF model is used to recognize the entities that is not defined in the vocabulary; The results of BMM and BiLSTM-CRF are sent to pattern recognizer which can generate new entities according to predefined pattern matching rules.

The algorithm is evaluated by recall, precision and F-1 score. We define $O = (O_1, \ldots, O_m)$ as the output entities of the algorithm, $G = (G_1, \ldots, G_m)$ as the entities manually annotated by physicians. The recall R = , precision P = $\frac{|O \cap G|}{|O|}$ and F1 = $\frac{2PR}{P+R}$.

3.3. Entity Normalization

In the medical records, there may exist different terms for a same entity. Thus, entity normalization is required to map the original term into a standard one, and further create entities by inheriting terms of the standard one.

Nine commonly used types of entity are defined in our KG, and each of them are described as below.

3.3.1. Disease

Disease is the core entity of our KG, and all the other entities are linked to disease. International classification of diseases, tenth revision (ICD-10) [30] is used as the standard of disease terms, and a controlled dictionary is used to map from original diagnosis term into an ICD-10 term. Corresponding disease entities are created for all normalized diagnosis diseases and their parent terms in ICD-10.

3.3.2. Gender

Gender has 2 values, i.e. male and female.

3.3.3. AgeRange

Age is a continuous variable, and hence it cannot serve as entity of the graph. To solve this issue, we quantize the age variable into 6 ranges, which are 0∼3, 3∼6, 6∼18, 18∼40, 40∼65 and 65 + .

3.3.4. Symptom

23146 symptoms are recognized by the entity recognition step from the chief complaint and present illness history of the medical data, and then they are normalized to 9767 standard symptoms using a mapping

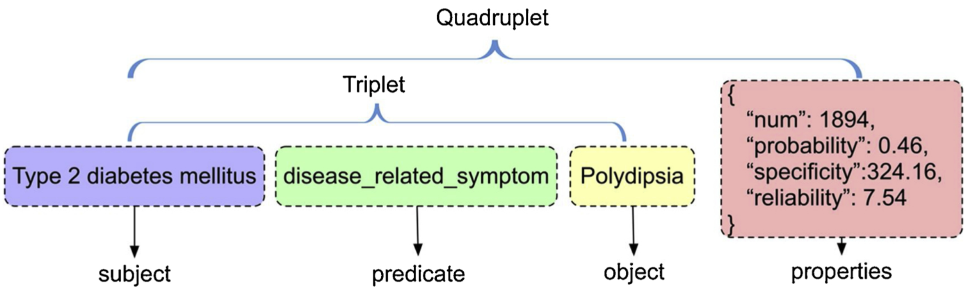


Fig. 2. An example of quadruplet.

dictionary. The items of the mapping dictionary were generated by NLP from EMRs and reviewed by medical experts.

3.3.5. *Exam*

The name of examination order is used as *Exam* entity name, which is due to the fact that the name for a same examination is unique in general for the same hospital.

3.3.6. *LabExam*

The name of laboratory examination order is used as *LabExam* entity name because of the similar reason for the Exam entity.

3.3.7. *LabItem*

Usually, a laboratory examination contains several examination items. In our KG, each item is regarded as a *LabItem* entity, and the item name is used as the entity name.

3.3.8. *Medicine*

Drug ingredient name is used as the medicine entity name, so that the drugs with the same ingredient will be regarded as the same entity in spite of different manufactures and/or forms.

3.3.9. *Surgery*

We use international classification of diseases, ninth revision (ICD-9) [31] as the standard of surgery terms, and a controlled dictionary is used to map original surgery name into an ICD-9 term.

3.4. *Relation Extraction*

In traditional KG, a fact is represented by a triplet, which refers to subject, predicate and object (SPO). Relation extraction aims to construct the triplet of two entities and their relation. In the medical domain, the disease is the central entity whether it is in case of diagnosis or treatment, only relationships between disease and other types of entities are needed in our applications. Therefore, 9 relations are defined in this paper. They are disease_related_gender, disease_related_agerange, disease_related_symptom, disease_related_exam, disease_related_labexam, disease_related_labitem, disease_related_medicine, disease_related_surgery, and disease_related_disease.

For each visit, there may be multiple diagnosis, especially for an inpatient visit. In our KG the normalized term and its parent terms in ICD-10 for all diagnosis diseases are created as disease entities. However, we just create the relations between main diagnosis entity and the other entities of the visit, in the KG. The underlying reason is that the main diagnosis disease is the purpose of the visit.

For example, if a main diagnosis disease of one visit is normalized to ICD code C16.902 (Gastric cancer NOS), then a list of triplets will be generated for each non-main diagnosis *d* in the visit, including (*C16.902, disease_related_disease, d*), (*C16.9, disease_related_disease, d*) and (*C16, disease_related_disease, d*). Similarly, for each prescribed medicine *m* in the visit, a list of triplets will be generated, i.e., (*C16.902, disease_related_medicine, m*), (*C16.9, disease_related_medicine, m*) and (*C16, disease_related_medicine, m*); for each recognized symptom *s* in this visit, a list of triplets will be generated, i.e., (*C16.902, disease_related_symptom, s*), (*C16.9, disease_related_symptom, s*) and (*C16, disease_related_symptom, s*). This applies to all the other types of entities.

3.5. *Property Calculation*

This subsection introduces property calculation for entities and relations, which is of critical importance for KG applications.

For each subject, two properties, i.e., the occurrence number, denoted by *No*, and the probability, should be calculated. Here, the occurrence number $N_o(S_i)$ is the number of visits whose main diagnosis is S_i, while the probability of S_i is defined as

$$\Pr(S_i) = \frac{N_o(S_i)}{\sum_{\ell=1}^{|S|} N_o(S_\ell)}, \tag{1}$$

where S denotes the set of diseases, and $|S|$ is the cardinality of S.

Based on the triplet structure, we introduce a *quadruplet* structure in view of the fact that the properties of the SPO are essential to rank the related entities, and are useful for applying the KG. Specifically, an additional field is defined to characterize the properties of the SPO. Fig. 2 illustrates a quadruplet to represent the fact that polydipsia is a related symptom with Type 2 diabetes mellitus.

For each relation, four basic properties are calculated, namely, co-occurrence number, probability, specificity and reliability.

The ***co-occurrence number*** $N_{co}(S_i, O_{ij})$ denotes the number of visits in which the subject S_i co-occurs with the object O_{ij} under the relation P.

The co-occurrence ***probability*** of a triplet measures the probability of the object given the subject, which is defined as.

$$\text{probability}_P(S_i, O_{ij}) \triangleq \Pr(O_{ij}\ |S_i) = \frac{N_{co}(S_i, O_{ij})}{N_o(S_i)}, \tag{2}$$

The ***specificity*** is defined as the ratio of the probability of O_{ij} given S_i to that under all subjects, which can be expressed as

$$\text{specificity }_P(S_i, O_{ij}) \triangleq \frac{\Pr(O_{ij}\ |S_i)}{\Pr(O_{ij}\ |S)}, \tag{3}$$

where $i = 1, \ldots, |S|$, $j = 1, \ldots, J_i$ with J_i being the number of the objects associated with the subject S_i, and

$$Pr(O_{ij}\ |S) = \frac{\sum_{\ell=1}^{|S|} N_{co}(S_\ell, O_{ij})}{\sum_{\ell=1}^{|S|} N_o(S_\ell)}. \tag{4}$$

Obviously, compared with the original probability, the value of specificity uses the relative value of probability to capture the significance of the object for the given subject.

In Table 1, we show an example of 5 related symptoms, probabilities and specificities of Parkinson disease (PD) and lung cancer, respectively. One can see that 3 of them, i.e., difficulties in turning over, tremor and bradykinesia, are more significant than insomnia and dizziness for PD diagnosis, although the dizziness is the most frequent symptom in terms of probability. Meanwhile, expectoration and bloody sputum are more significant than shortness of breath, chest pain and cough.

Table 1
Specificity of symptoms for Parkinson's Disease and Lung Cancer

	Symptom		Probability	Specificity
	English	*Chinese*		
Parkinson Disease 帕金森病	Dizziness	头昏	0.2579	2.8479
	Tremor	震颤	0.2293	**8.5357**
	Insomnia	失眠	0.2029	2.6009
	Bradykinesia	动作缓慢	0.1636	**8.6692**
	Difficulty in turning over	翻身困难	0.1235	**7.8383**
Lung cancer 肺恶性肿瘤	Cough	咳嗽	0.0939	2.7627
	Expectoration	咳痰	0.0633	**3.0085**
	Shortness of breath	气促	0.0395	2.0369
	Chest pain	胸痛	0.0255	2.2157
	Bloody sputum	痰中带血	0.0246	**4.0259**

The ***reliability*** is defined by

$$\text{reliability}_P(S_i,\ O_{ij}) = log_{10}(\max(1,\ 1 + N_{\text{co}}(S_i,\ O_{ij}) - N_{co}^{min})) + R \tag{5}$$

It is worth noting that additional properties can be readily defined and integrated into our proposed quadruplet. Here, let us take the relation disease_related_labitem as an example. An abnormal laboratory test result often indicates the onset of some diseases. Hence, it is important to calculate the rate of test abnormality. In our study, the laboratory test result is classified into three categories, which are normal, high, and low, respectively. Then the number and rate of each category are defined as extra properties for disease_related_labitem. These properties will be used for LabItem ranking and CDSS applications.

3.6. Graph Cleaning

Graph cleaning is a key step against the impact due to the data noise. In order to remove the invalid entities and relations, some primary cleaning rules are adopted in our study, including

- entity deletion, where an entity S_i should be deleted from the graph if the occurrence number is less than a specified threshold, i.e., $N_o(S_i) \leq N_o^{\min}$;
- quadruplet deletion, where a quadruplet should be deleted if the co-occurrence number of $(S_i,\ O_{ij})$ is less than a given threshold, i.e., $N_{\text{co}}(S_i,\ O_{ij}) \leq N_{\text{co}}^{\min}$, or its probability is less than a certain threshold, i.e., $\text{probability}_p(S_i,\ O_{ij}) \leq P_{ij}^{\min}$.

In our study, both $N_o^{\min}$ and $N_{\text{co}}^{\min}$ are set to be 10, while $P_{ij}^{\min}$ equals to 0.01. In practice, all thresholds could be adjusted for different relations and hospitals.

3.7. Related-entity Ranking

For a medical knowledge graph, a disease may have multiple symptoms, and it also corresponds to multiple medicines in general. Therefore, it cannot be directly applied for recommendation of symptom, examination, medicine, etc. in CDSS. To address this issue, we should perform the subject related entities ranking.

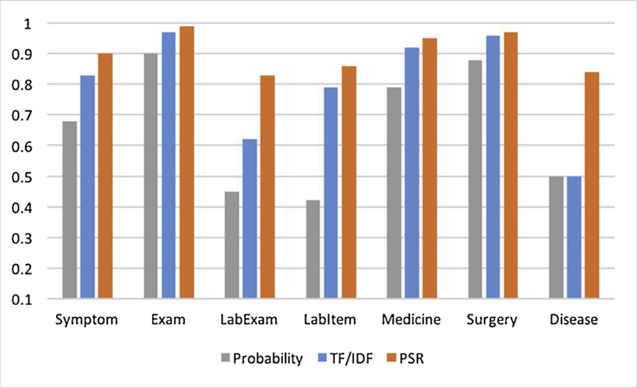


Fig. 3. Performance comparison of ranking functions by NDCG@10. For the 7 relations, NDCG@10 by probability are 0.68, 0.9, 0.45, 0.42, 0.79, 0.88, 0.5, with average value 0.660; NDCG@10 by TF/IDF are 0.83, 0.97, 0.62, 0.79, 0.92, 0.96, 0.5, with average value 0.799; NDCG@10 by PSR are 0.9, 0.99, 0.83, 0.86, 0.95, 0.97, 0.84, with average value 0.906.

Towards this end, we first apply the TF-IDF as a score function of $(S_i,\ O_{ij})$ under the relation P, which can be expressed as

$$\text{TFIDF}_P(S_i,\ O_{ij}) \triangleq \text{TF}_P(S_i,\ O_{ij})\cdot\text{IDF}_P(O_{ij}), \tag{6}$$

where $\text{TF}_P(S_i,\ O_{ij}) \triangleq \text{probability}_p(S_i,\ O_{ij})$, $\text{IDF}_P(O_{ij}) = \log\left(\frac{|\mathscr{S}|}{|\mathscr{S}_{ij}|}\right)$, and the set $\mathscr{S}_{ij} \triangleq \{S_m \in \mathscr{S}O_{mn} = O_{ij},\ \forall\ m,\ n\}$, for all i and j. In simple words, $\mathscr{S}_{ij}$ is the set of subjects who is related with O_{ij}.

In the medical domain, when we mention the relationship strength between two entities, both the conditional probability and specificity are considered. Besides, since the two values are statistics-based, the reliability of the values needs to be considered as well. Hence, a novel score function, termed as *Probability-Specificity-Reliability (PSR)* is proposed, which is defined as

$$\text{PSR}_P(S_i,\ O_{ij}) = \text{probability}_p(S_i,\ O_{ij})\cdot\text{specificity}_p(S_i,\ O_{ij})\cdot\text{reliability}_p(S_i,\ O_{ij}). \tag{7}$$

PSR function is used to rank and retrieve the most related entities for a given subject, from all its related entities extracted in the relation extraction step.

Specifically, for the relation disease_related_labitem, we propose an enhanced PSR function, as detailed below,

$$\text{PSR}_{disease\_related\_labitem}(S_i,\ O_{ij}) = \text{PSR}_P(S_i,\ O_{ij})\cdot\text{RAR}(S_i,\ O_{ij}), \tag{8}$$

with $\text{RAR}(S_i,\ O_{ij}) \triangleq \frac{\text{AR}(S_i, O_{ij})}{\text{AR}(\widetilde{S}_i, O_{ij})}$, $\text{AR}(S_i,\ O_{ij}) \triangleq \frac{N_{abn}(S_i, O_{ij})}{N_{co}(S_i, O_{ij})}$, and $\text{AR}(\tilde{S}_i, O_{ij}) \triangleq \frac{\sum_{\ell \neq i} N_{abn}(S_\ell, O_{\ell j})}{\sum_{\ell \neq i} N_{\text{co}}(S_\ell, O_{\ell j})}$, where $N_{abn}(S_i,\ O_{ij})$ denotes number of visits in which S_i co-occurs with O_{ij} and the test result of O_{ij} is abnormal.

Note that

- $\text{AR}(S_i,\ O_{ij})$ represents the abnormal result rate of lab item O_{ij} to disease S_i;
- $AR(\widetilde{S}_i,\ O_{ij})$ represents the abnormal result rate of lab item O_{ij} to all

Table 2
Meaning of different label rates.

Relation	***	**	*	-
disease_related_symptom	Very useful for disease diagnosis	Useful for disease diagnosis	related but not main symptoms for diagnosis	Not related
disease_related_medicine	Main medicine for the disease	NA	Auxiliary medicine for the disease	Not related
disease_related_labExam	Closely related with the disease	NA	Related with the disease	Not related
disease_related_exam	Closely related with the disease	NA	Related with the disease	Not related
disease_related_surgery	Main surgery for the disease	NA	Auxiliary surgery for the disease	Not related
disease_related_labItem	Closely related with the disease	NA	NA	Not related
disease_related_disease	Complication/cause of the disease	NA	Comorbidity	Not related

Table 3
Comparison of disease_related_medicine ranking

	Probability	TF/IDF	PSR
Related symptoms of Lung Cancer	cough 咳嗽(**)	cough 咳嗽(**)	cough 咳嗽(**)
	expectoration 咳痰(**)	expectoration 咳痰(**)	expectoration 咳痰(**)
	shortness of breath 气促(**)	bloody sputum 痰中带血(***)	bloody sputum 痰中带血(***)
	insomnia 失眠(-)	shortness of breath 气促(**)	dry cough 干咳(**)
	chest pain 胸痛(**)	chest pain 胸痛(**)	shortness of breath 气促(**)
	bloody sputum 痰中带血(***)	hemoptysis 咯血(***)	hemoptysis 咯血(***)
	loss of appetite 食欲不振(-)	dry cough 干咳(**)	chest pain 胸痛(**)
	weight loss 体重下降(**)	heart tired 心累(-)	weight loss 体重下降(**)
	mental deficiency 精神欠佳(-)	hoarseness 声音嘶哑(**)	chest and back pain 胸背部疼痛(**)
	dry cough 干咳(**)	fever 发热(**)	night sweat 盗汗(**)
NDCG@10	0.59	0.80	0.85
Related laboratory items of Lung Cancer	lymphocyte percentage 淋巴细胞百分比(-)	cytokeratin 19 fragment 细胞角蛋白19片段(***)	carcinoembryonic antigen 癌胚抗原(***)
	lymphocyte count 淋巴细胞计数(-)	fibrinogen 纤维蛋白原(-)	cytokeratin 19 fragment 细胞角蛋白19片段(***)
	monocyte count 单核细胞计数(-)	HIV type 1 RNA 人类免疫缺陷病毒1型RNA(-)	neuron-specific enolase 神经元特异性烯醇化酶(***)
	neutrophil percentage 中性粒细胞百分比(-)	carcinoembryonic antigen 癌胚抗原(***)	CA242 糖类抗原CA242(***)
	monocyte perfentage 单核细胞百分比(-)	hepatitis c virus rna 丙型肝炎病毒RNA(-)	CA125 糖类抗原CA125(***)
	platelet count 血小板计数(-)	neuron-specific enolase 神经元特异性烯醇化酶(***)	fibrinogen 纤维蛋白原(-)
	neutrophil count 中性粒细胞计数(-)	ca125 糖类抗原CA125(***)	CA19-9 糖类抗原CA19-9(-)
	white blood cell count 白细胞计数(-)	CD3 CD4 Assisted/Induced T Lymphocyte CD3 + CD4 + 辅助/诱导T淋巴细胞(-)	human chorionic gonadotropin 人绒毛膜促性腺激素(-)
	eosinophils percentage 嗜酸性粒细胞百分比(-)	hepatitis b virus dna 乙型肝炎病毒DNA(-)	CD3 CD4/CD3 CD8 cell ratio CD3 + CD4 + /CD3 + CD8 + 细胞比值(-)
	hematocrit 红细胞比容(-)	platelet hematocrit 血小板比容(-)	CD3 CD4 Assisted/Induced T Lymphocyte CD3 + CD4 + 辅助/诱导T淋巴细胞(-)
NDCG@10	0.00	0.72	1.00

other diseases than S_i;

- RAR(S_i, O_{ij}) represents the relative abnormal rate of lab item O_{ij} to disease S_i comparing with average abnormal rate to all other diseases. The higher value, the more significant is O_{ij} with S_i.

The enhanced PSR function can reflect the correlation between laboratory result abnormality and disease. Laplace smoothing was used to avoid $\sum_{\ell \neq i} N_{\mathrm{co}}(S_\ell, O_{\ell j})$ or $AR(\tilde{S}_i, O_{ij})$ equals with 0.

3.8. Graph Embedding

In this paper, PrTransH [21] is used to learn embedding vectors from the constructed quadruplet-based medical KG. As PrTransH can learn the probability of one fact into the embedding vectors, by considering the probability in the training objective. As a result, embedding vectors are learned for all entities and relations in the medical KG. The vector dimension is set to 100 in our experiment.

4. RESULTS

4.1. Entity Recognition

We compared the performance of single BiLSTM-CRF and the proposed hybrid model on symptom recognition, on a dataset which is composed of 1000 present illness history and their labeled symptoms by physicians. The recall, precision and F1-score of the single BiLSTM-CRF are 0.9368, 0.9482 and 0.9425. The results of the hybrid model are improved to 0.9689, 0.9727 and 0.9708, respectively.

4.2. Related-entity Ranking

A data-set is built to evaluate the performance of the proposed PSR function. For each disease, all original related entities of the 7 relations, i.e., disease_related_symptom, disease_related_exam, disease_related_labexam, disease_related_labitem, disease_related_medicine, disease_related_surgery and disease_related_disease are labeled by physicians with different rates as ground truth according to clinical guideline and medical literature. The meaning of different label rates is illustrated in Table 2. As manually labeling is quite labor-intensive, 10 diseases are randomly selected.

We compare the ranking performance between TF/IDF in (6) and PSR in (7) with ranking by single probability, for the 7 relations respectively. Normalized discounted cumulative gain (NDCG) [32] is used as the measure since it is designed to evaluate ranking multiple rates items. From Fig. 3, we can see that the proposed PSR function outperforms both the probability and TF/IDF based functions for all relations.

Top 10 ranking results of relations disease_related_symptom and disease_related_labitem of Lung Cancer (C34.901) is shown in Table 3. We can see that PSR can find the most significant symptoms and tumor markers for Lung Cancer.

4.3. Graph Embedding

The embedding vectors of one entity learned by PrTransH preserve its structural relationships with different types of entities. Therefore, if two diseases have similar related entities, their embedding vectors are similar. Based on this principle, we cluster the diseases using the trained embedding vectors in order to validate its effectiveness. Firstly,

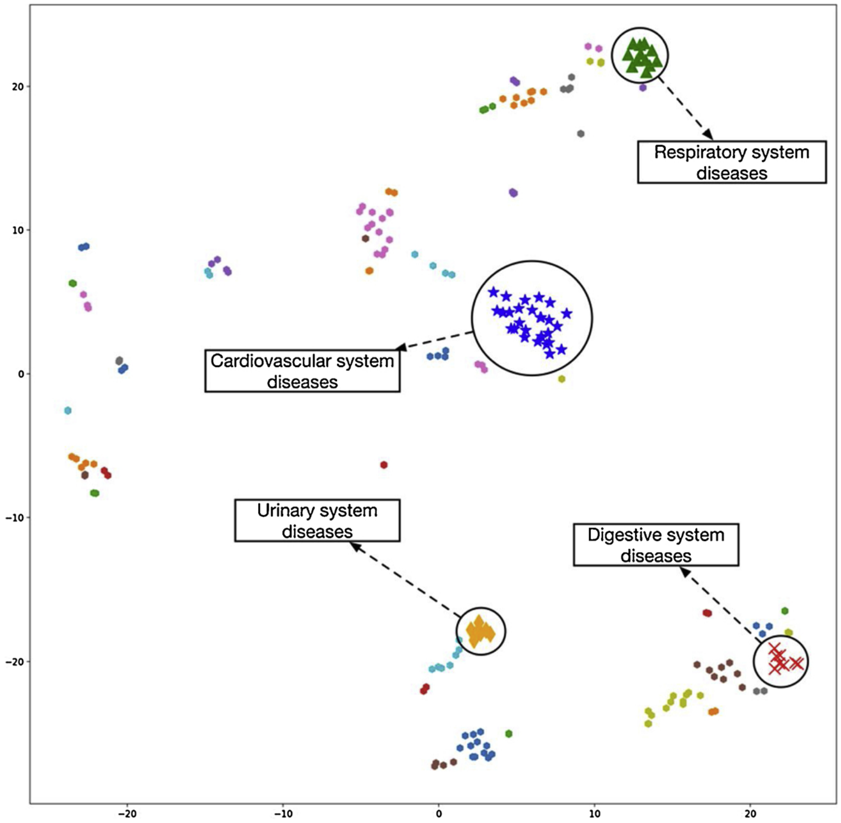


Fig. 4. Clusters of disease embedding vectors.

Table 4
Numbers of entities and quadruplets in the constructed KG

Entity types	Entity numbers	Relations	Quadruplet numbers
Disease	7,465	disease_related_disease	18876
Gender	2	disease_related_gender	12,094
AgeRange	6	disease_related_agerange	9,314
Symptom	9,769	disease_related_symptom	31,933
Medicine	1,588	disease_related_medicine	68,719
Surgery	1,540	disease_related_surgery	14,210
Exam	219	disease_related_exam	10,276
LabExam	1,121	disease_related_labexam	39,987
LabItem	798	disease_related_labitem	373,685
Total	**22,508**	**Total**	**579,094**

we cluster the top 500 occurred diseases in embedding space using DBSCAN [33], then project the embedding vectors to 2D space, and visualize clusters of diseases. As shown in Fig. 4, the diseases are grouped into different clusters, and diseases of the same physiological system are grouped together, which proves that the learned graph embedding vector is good semantic representation of diseases. To illustrate, four typical clusters are circled and annotated with disease category (Supplementary Table 1).

4.4. Constructed medical KG

The constructed medical knowledge graph contains 9 entity types, totally 22,508 entities and 579,094 quadruplets. The quadruplets cover relations between disease and all 9 entity types. The numbers of entities and quadruplets could be found in Table 4.

To visualize the constructed KG, we developed a web application in hospital intranet. Fig. 5 is the screenshot of the KG of Lung Cancer (肺恶性肿瘤), the central node indicates entity of Lung Cancer, and its linked nodes are related entities of Lung Cancer. Moreover, each edge in the graph corresponds with a quadruplet of the KG. The main properties of the entity or quadruplet will be shown as tips, while the mouse focus on the node or edge. The other properties are not shown in the tips as they are difficult to understand by physicians.

5. APPLICATIONS

The constructed knowledge graph can be applied in many medical problems. In this paper we demonstrated the knowledge graph to three typical problems: clinical decision support system, medical information retrieval and knowledge transferring with neural networks.

5.1. Clinical decision support system

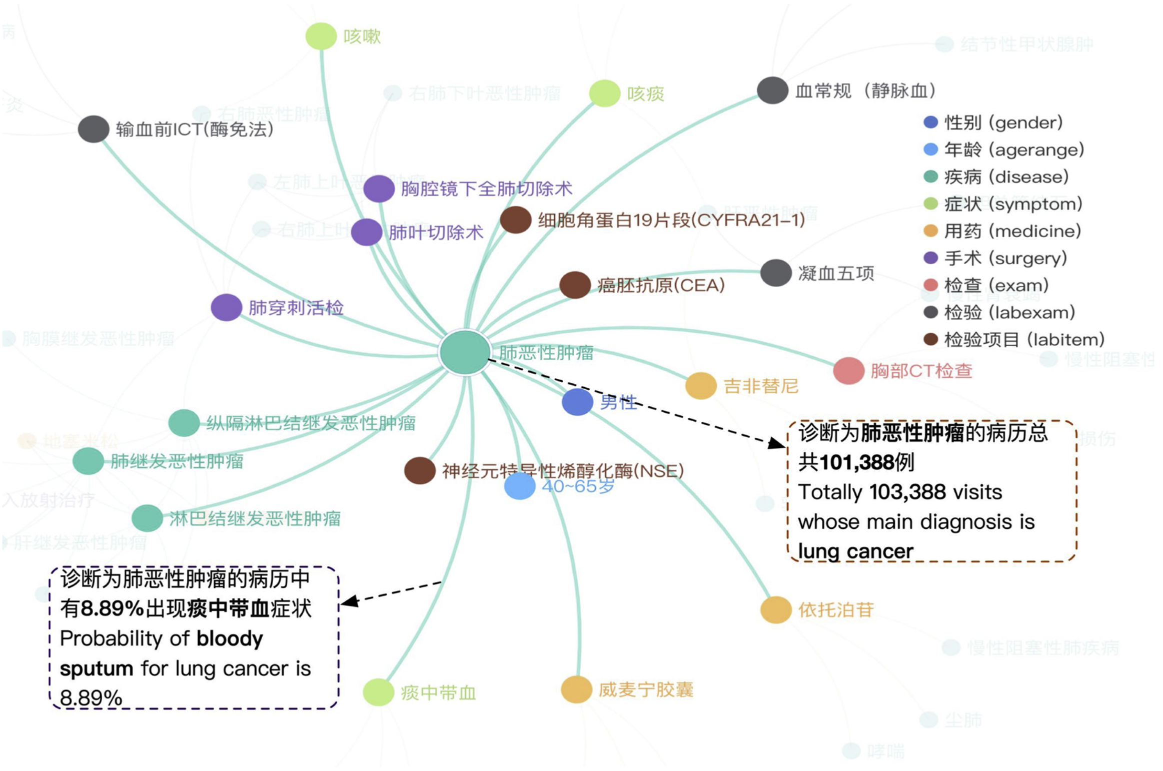
CDSS is designed to recommend possible diagnosis and treatment given patient age, gender, symptoms, laboratory item results and other factors. Take diagnosis decision support as example, according to Naïve Bayes, the probability of the patient's disease is D_i can be calculated by the following formula:

$$Pr(D_i Age, Gender,\ S_1 ... S_m,\ L_1 ... L_n) \tag{9}$$

$$= \frac{1}{J} \times P(D_i) \times P(Age\ |D_i) \times P(Gender\ |D_i) \times \prod_{t=1}^{m} P(S_t\ |D_i) \times \prod_{j=1}^{n} P(L_j\ |D_i) \tag{10}$$

where $\mathscr{J} = Pr(\, Age, Gender, S_1 ... S_m,\ L_1 ... L_n)$, $S_{1,...,}S_m$ indicate m symptoms of the patient and $L_1,..., L_n$ indicates n laboratory items of the patient.

To rank the possible diseases, the joint distribution $\mathscr{J}$ can be ignored as it is the same for all D_i. $P(D_i)$ is the prior probability of D_i which is a statistical property of entity D_i. $P(Age|D_i)$ is the probability included in the properties field of quadruplet {D_i, disease_related_agerange, Age, properties}. It is the same as calculating the gender conditional probability $P(Gender|D_i)$ and symptom conditional probability $P(S_t|D_i)$. The conditional probability of result of laboratory item L_j is abnormal for



诊断为**肺恶性肿瘤**的病历总共**101,388**例
Totally **103,388** visits whose main diagnosis is lung cancer

诊断为肺恶性肿瘤的病历中有**8.89%**出现**痰中带血**症状
Probability of **bloody sputum** for lung cancer is 8.89%

Fig. 5. Real-world data KG of Lung Cancer.

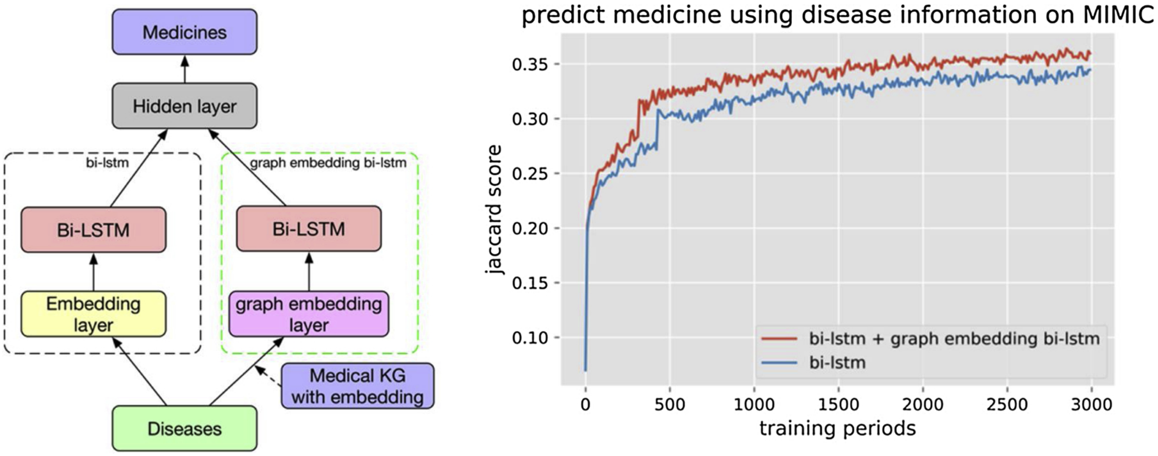


Fig. 6. Transferring graph embedding to neural network and its result.

disease D_i is equals with the abnormal rate which is included in the properties field of quadruplet {D_i, disease_related_labitem, L_j, properties}. If the quadruplet does not exist in the KG, the conditional probability is set to 0. Finally, the most possible diagnosis could be deduced.

Without loss of generality, the medical KG also be used to recommend medical order (exam, laboratory exam and medicine) to physician for a patient.

The CDSS system is developed and being used at clinical department of Southwest Hospital. We evaluated the performance of the KG based diagnosis recommendation. 5000 visits data that are not used for KG construction were selected. The data covered the most frequently happened 500 diseases. For each visit, the main diagnosis given by the physician is regarded as the ground truth, the top-5 and top-10 recall of the diagnosis recommendation model is 74.20% and 88.76%, respectively.

5.2. Information retrieval

An important application of DPAP platform of the hospital is to provide second-level searching on medical data, like a Google/Baidu in hospital intranet, thanks to the platform doctors can retrieve the interested medical records very quickly. However, there are many medicines and examination results in a visit, especially for inpatient visit, doctors need to look through the medical records one by one in order to find out the key medicines and abnormal examination results, which is time consuming. In order to overcome the problem, the KG was used to rank related medicines and examination records within a visit data. Specially, these records are sorted by its ranking score related to the main diagnosis from the KG, and thus the doctors can quickly spot the key information by simply looking at the first few records.

To evaluate the performance, we randomly choose 1000 inpatient visits whose main diagnosis disease is in the scope of labeled diseases of evaluation data-set in related-entity ranking step. For each of the inpatient visits, if the labeled entities really occur during this visit, it is regarded as related entity. Based on the data set, NDCG@5 [32] is used as measure since the top 5 most related entities is shown by default in the product, the average result is 0.87.

5.3. Knowledge transferring with neural network

To demonstrate the application of the learned graph embedding vector, this paper performed an experiment applying graph embedding to a neural network task. The task is predicting medicine prescription by diseases using the MIMIC3 dataset [34]. The experiment is performed using TensorFlow. The task was originally trained by a single Bi-LSTM network, the Jaccard score between predicted medicines and actual medicines was 34.7%. As shown in Fig. 6, another Bi-LSTM component using the graph embedding was added to the original network, then the Jaccard score rise to 36.4%. In addition, the new model can converge more quickly.

The embedding vectors were fixed and only Bi-LSTM parameters were learned during the training process. Meanwhile, all the hyper parameters use the default values. The experiment shows that performance improvement can be obtained to general neural network task by taking the medical graph embedding through Bi-LSTM component.

6. CONCLUSIONS AND DISCUSSIONS

This paper establishes a systematic procedure to construct a quadruplet-based medical KG from large-scale EMRs. The evaluation result shows that the constructed KG is high-quality, and the KG is applied successfully due to the real-world statistical properties of the quadruplet. The proposed ranking function PSR outperforms other algorithms under all relations, and the disease clustering results validate the efficacy of the learned embedding vector as entity's semantic representation.

It is proved that the KG can be used with Naïve Bayes to make inference of possible diseases and recommend medical orders. It also can be applied in medical records search engine to rank related results. In addition, transferring the knowledge through embedding vector into neural network can make it achieve better performance. Other than the above demonstrated applications, the knowledge graph could be used in many other directions, including unknown knowledge exploration for clinical research, building self-diagnosis utilities and health question answering systems for patients, medical record quality control, etc.

It would be an interesting and challenging direction to combine the real-world medical KG with that built from medical textbooks and literatures. The former one has statistically properties for each fact which makes it easier to be applied. Meanwhile, the latter one has less noise and the knowledge is more acceptable by physicians.

FUNDING

This work is supported in part by the National Key Research Program under Grant 2018YFC0116700.

CONTRIBUTORS

LL, PW, JY and JJ were responsible for the design of the systematic procedure. LL, PW, YW, SL, JJ, SW and YL were responsible for data processing, algorithms and experiments. KG applications are developed and evaluated by LL, YW, SL and JJ.

LL wrote the first draft of the manuscript, and PW, JY, ZS, BT, THC, SW and YL made the revisions to it.

All authors, LL, PW, JY, YW, SL, JJ, ZS, BT, THC, SW and YL approved the version of the manuscript to be published, and agreed to be accountable for all aspects of the work.

Declaration of Competing Interest

None

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.artmed.2020.101817.

References

[1] Zhao Z, Han S, So I. Architecture of knowledge graph construction techniques. International Journal of Pure and Applied Mathematics 2018;118(19):1869–83.

[2] Barnett GO, Cimino JJ, Hupp JA, et al. DXplain. An evolving diagnostic decision-support system. Jama 1987;258(1):67–74.

[3] Bisson LJ, Komm JT, Bernas GA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. The American journal of sports medicine 2014;42(10):2371–6.

[4] Miller RA. Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary. Journal of the American Medical Informatics Association 1994;1(1):8–27.

[5] Wang M, Liu M, Liu J, et al. Safe medicine recommendation via medical knowledge graph embedding. arXiv preprint arXiv:1710.05980.2017.

[6] Tang H, Ng JHK. Googling for a diagnosis–use of Google as a diagnostic aid: internet based study. BMJ 2006;333(7579):1143–5.

[7] Gann B. Giving patients choice and control: health informatics on the patient journey. Yearbook of medical informatics 2012;21(01):70–3.

[8] hwe MA, Middleton B, Heckerman DE, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. Methods of information in Medicine 1991;30(04):241–55.

[9] Kovačević A, Dehghan A, Filannino M, et al. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. Journal of the American Medical Informatics Association 2013;20(5):859–66.

[10] Tang B, Cao H, Wu Y, et al. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC Medical Informatics and Decision Making 2013;13(1):1–10.

[11] Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics 2017;34(8):1381–8.

[12] Zhang Y, Wang X, Hou Z, et al. Clinical Named Entity Recognition From Chinese Electronic Health Records via Machine Learning Methods. JMIR medical informatics 2018;6(4):e50.

[13] Ji B, Liu R, Li S, et al. A hybrid approach for named entity recognition in Chinese electronic medical record. BMC medical informatics and decision making 2019;19(2):64.

[14] Li H, Chen Q, Tang B, et al. CNN-based ranking for biomedical entity normalization. BMC bioinformatics 2017;18(11):385.

[15] Lou Y, Zhang Y, Qian T, et al. A transition-based joint model for disease named entity recognition and normalization. Bioinformatics 2017;33(15):2363–71.

[16] Li F, Jin Y, Liu W, et al. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. JMIR medical informatics 2019;7(3):e14830.

[17] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information processing & management 1988;24(5):513–23.

[18] Vechtomova O, Robertson SE. A domain-independent approach to finding related entities. Information Processing & Management 2012;48(4):654–70.

[19] Kang C, Yin D, Zhang R, et al. Learning to rank related entities in Web search. Neurocomputing 2015;166:309–18.

[20] Wang M, Liu M, Liu J, et al. Safe medicine recommendation via medical knowledge graph embedding. arXiv preprint arXiv 2017. 1710.05980.

[21] Li L, Wang P, Wang Y, et al. PrTransH: Embedding Probabilistic Medical Knowledge from Real World EMR Data. arXiv preprint arXiv 2019. 1909.00672.

[22] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems 2013:2787–95.

[23] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes. Twenty-Eighth AAAI conference on artificial intelligence 2014.

[24] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion. Twenty-ninth AAAI conference on artificial intelligence 2015.

[25] Finlayson SG, LePendu P, Shah NH. Building the graph of medicine from millions of clinical narratives. Scientific data 2014;1:140032.

[26] Rotmensch M, Halpern Y, Tlimat A, et al. Learning a health knowledge graph from electronic medical records. Scientific reports 2017;7(1):5994.

[27] Zhao C, Jiang J, Xu Z, et al. A study of EMR-based medical knowledge network and its applications. Computer methods and programs in biomedicine 2017;143:13–23.

[28] Zhao C, Jiang J, Guan Y, et al. EMR-based medical knowledge representation and inference via Markov random fields and distributed representation learning. Artificial intelligence in medicine 2018;87:49–59.

[29] Shen Y, Zhang L, Zhang J, et al. CBN: Constructing a clinical Bayesian network based on data from the electronic medical record. Journal of biomedical informatics 2018;88:1–10.

[30] Bramer GR. International statistical classification of diseases and related health problems. Tenth revision. World Health Stat Q 1988;41:32–6.

[31] Slee VN. The International classification of diseases: ninth revision (ICD-9). Annals of internal medicine 1978;88(3):424–6.

[32] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 2002;20(4):422–46.

[33] Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of Second International Conference on Knowledge Discovery and Data Mining 1996.

[34] Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Scientific data 2016;3:160035.