Original Paper

Head and Tail Entity Fusion Model in Medical Knowledge Graph Construction: Case Study for Pituitary Adenoma

An Fang[1,2], MS; Pei Lou[2], MS; Jiahui Hu[2], PhD; Wanqing Zhao[2], MSc; Ming Feng[3], MD; Huiling Ren[2], MSL; Xianlai Chen[4,5], PhD

[1]Life Science College, Central South University, Changsha, China

[2]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

[3]Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China

[4]Big Data Institute, Central South University, Changsha, China

[5]National Engineering Lab for Medical Big Data Application Technology, Central South University, Changsha, China

Corresponding Author:
Xianlai Chen, PhD
Big Data Institute
Central South University
932 South Lushan Road
Changsha, 410083
China
Phone: 86 731 88879583
Email: chenxianlai@csu.edu.cn

Abstract

Background: Pituitary adenoma is one of the most common central nervous system tumors. The diagnosis and treatment of pituitary adenoma remain very difficult. Misdiagnosis and recurrence often occur, and experienced neurosurgeons are in serious shortage. A knowledge graph can help interns quickly understand the medical knowledge related to pituitary tumor.

Objective: The aim of this study was to develop a data fusion method suitable for medical data using data of pituitary adenomas integrated from different sources. The overall goal was to construct a knowledge graph for pituitary adenoma (KGPA) to be used for knowledge discovery.

Methods: A complete framework suitable for the construction of a medical knowledge graph was developed, which was used to build the KGPA. The schema of the KGPA was manually constructed. Information of pituitary adenoma was automatically extracted from Chinese electronic medical records (CEMRs) and medical websites through a conditional random field model and newly designed web wrappers. An entity fusion method is proposed based on the head-and-tail entity fusion model to fuse the data from heterogeneous sources.

Results: Data were extracted from 300 CEMRs of pituitary adenoma and 4 health portals. Entity fusion was carried out using the proposed data fusion model. The F1 scores of the head and tail entity fusions were 97.32% and 98.57%, respectively. Triples from the constructed KGPA were selected for evaluation, demonstrating 95.4% accuracy.

Conclusions: This paper introduces an approach to fuse triples extracted from heterogeneous data sources, which can be used to build a knowledge graph. The evaluation results showed that the data in the KGPA are of high quality. The constructed KGPA can help physicians in clinical practice.

(JMIR Med Inform 2021;9(7):e28218) doi: 10.2196/28218

KEYWORDS

knowledge graph; pituitary adenoma; entity fusion; similarity calculation

Introduction

Pituitary adenoma is one of the most common central nervous system tumors. Most of the benign adenomas are characterized by swelling growth, which can be cured by surgery or medicine [1]. However, a small number of pituitary adenomas are not sensitive to surgery, radiotherapy, and drug therapy, and metastasis will lead to pituitary adenocarcinoma [2]. At present, there are difficulties in the diagnosis and treatment of pituitary adenoma [3]. In some cases, pituitary adenocarcinoma can even be life-threatening [4] and the prognosis is extremely poor. Therefore, pituitary adenoma has become a hot topic in life science research, and an open knowledgebase of pituitary adenoma is needed.

A knowledge graph is a general framework for formal description of knowledge, which can describe knowledge in the form of triples as a “head entity-relation-tail entity,” one of the most popular knowledge representation methods currently adopted [5]. Well-known open-domain knowledge graphs include Freebase, DBpedia, YAGO, and NELL, among others [6]. Knowledge graphs are also widely used in the medical field. Gong et al [7] proposed a method to build a diabetes knowledgebase by mining the web; they extracted knowledge from the semistructured content of the vertical portal and then mapped the information onto a unified knowledge graph. Ernst et al [8] constructed a biomedical science knowledge graph in which they extracted data using distant supervision methods and used logical reasoning for consistency checks. Rotmensch et al [9] designed an automatic extraction framework to directly extract diseases and symptoms from electronic medical records (EMRs), and automatically constructed a knowledge graph.

Data fusion is an important step of the integration of heterogeneous data in the construction of knowledge graphs. Entity fusion includes methods based on character similarity, clustering, deep learning, and others. Zhang et al [10] proposed a novel multisource medical data integration and mining solution for better health care services, which can search for similar medical records in a time-efficient and privacy-preserving manner. Wang et al [11] extracted different semantic words using multimodal trees and performed multigranularity feature fusion on the data. Li et al [12] proposed a novel fusion-embedding learning model, G2SKGE, which aims to learn the subgraph structure information of the entity in a knowledge graph. Li et al [13] proposed an approach to build a knowledge graph for hepatocellular carcinoma, and applied a biomedical information extraction system to filter and fuse the data.

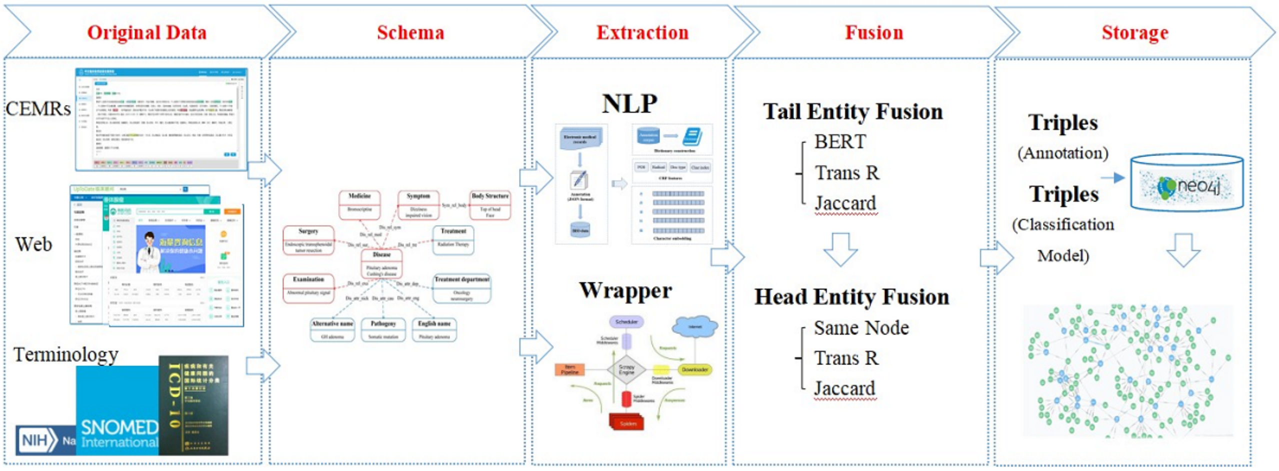
In this study, we extracted data from patient EMRs and medical websites, fused the entities using our proposed head-and-tail entity fusion model, and constructed a medical knowledge graph for pituitary adenoma (KGPA). The main contributions of this study are as follows. First, there is currently no Chinese knowledgebase for pituitary adenoma. Therefore, this study presents the complete process of knowledge graph construction, which was used to construct the KGPA. Second, to integrate the data extracted from different sources, we propose a fusion method suitable for medical data that was used in the process of KGPA construction. The method includes two steps: tail entity fusion and head entity fusion. Finally, knowledge of pituitary adenoma, such as the typical symptoms of different pituitary adenoma–related diseases, can be clearly revealed by searching the KGPA. According to doctors’ feedback on use of the KGPA, the content displayed in the KGPA was considered to be consistent with the actual clinical situation.

Methods

Overview

According to the characteristics of pituitary adenoma diseases combined with the characteristics of Chinese electronic medical records (CEMRs) and Chinese health websites, we designed the construction framework of the KGPA, as shown in Figure 1, which includes 5 steps: raw data collection, schema design, data extraction, data fusion, and data storage and visualization. Each step is introduced in detail below, with emphasis on the proposed data fusion model.

Figure 1. Process for construction of the knowledge graph for pituitary adenoma. CEMR: Chinese electronic medical record; NLP: natural language processing; BERT: bidirectional encoder representations from transformer.



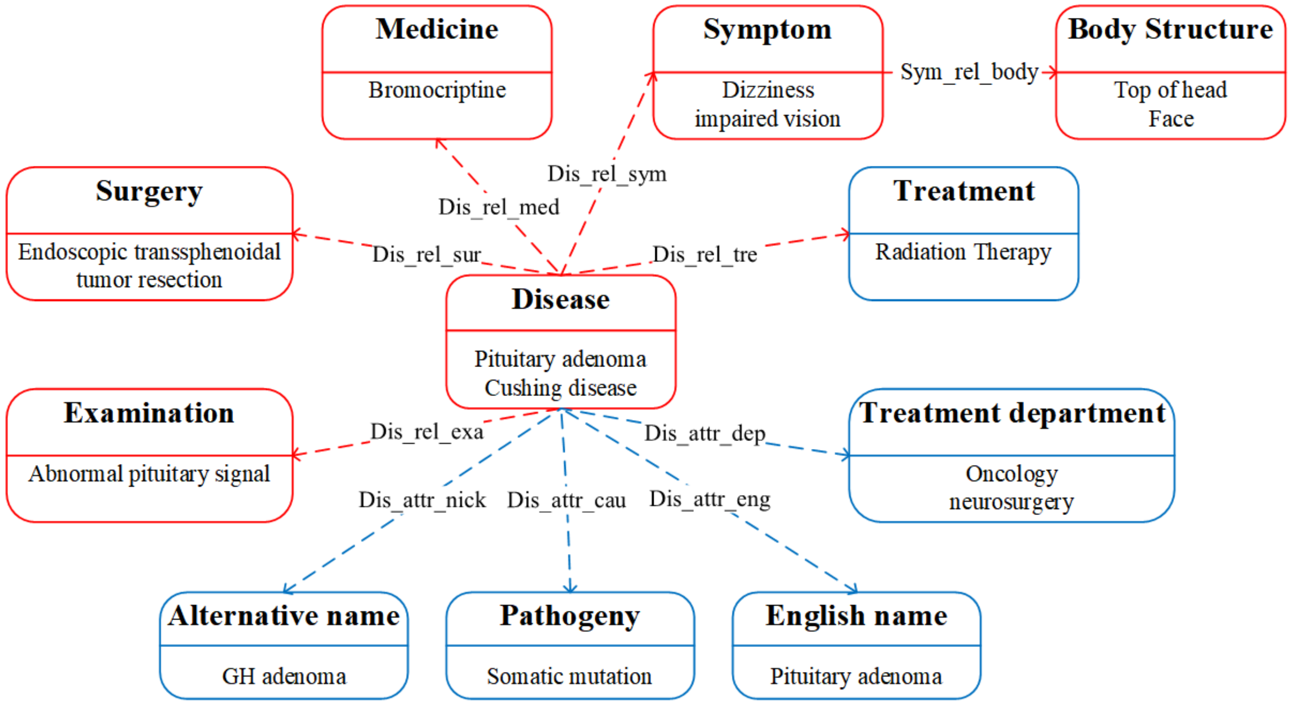
Data Schema

The knowledge graph includes a data layer and a schema layer [14]. Entities, relations, and attributes in the data layer are regulated and restricted by the schema. The schema was based on several open-access authoritative terminologies and ontologies, including the UMLS Semantic Network [15], the concept definitions in SNOMED-CT [16], and the International Statistical Classification of Diseases and Related Health Problems (ICD-10). In addition, the natural language processing datasets defined by the Informatics for Integrating Biology & the Bedside [17] and CEMRs Entity and Relations Annotation

Specifications defined by Harbin Institute of Technology [18] were also referenced for this task. With the help of clinical experts, a combination of top-down and bottom-up approaches was used to construct the KGPA schema.

In our previous study of CEMRs data extraction, we found that the medical diagnosis and treatment activities could be summarized based on symptoms (symptom) and abnormal results (examination) [19]. The doctor will give a comprehensive diagnosis conclusion (disease) and corresponding treatment measures (surgery, medicine). Therefore, the mentioned entities and the relations between them were abstracted for design of the schema. The CEMRs are detailed but contain a limited number of concepts; therefore, we extracted data from medical websites to expand the concepts. Through analyzing the data types of the websites, six types of concepts were added to the schema: pathogeny, treatment, examination, treatment department, English name, and alternative name. The most frequently used disease term in websites was selected as the concept of the disease, and then treatment and examination were defined as related entities. Pathogeny, treatment department, English name, and alternative name were defined as the attributes of the disease. Attributes can be used to describe the internal characteristics of the disease entities; the more attributes there are, the more complete the information of the entity will be [20]. The KGPA schema is shown in Figure 2.

Figure 2. Schema of the knowledge graph for pituitary adenoma (KGPA). Concepts extracted from Chinese electronic medical records are in red. Concepts extracted from health websites are in blue. GH: growth hormone.



Data Extraction

Process

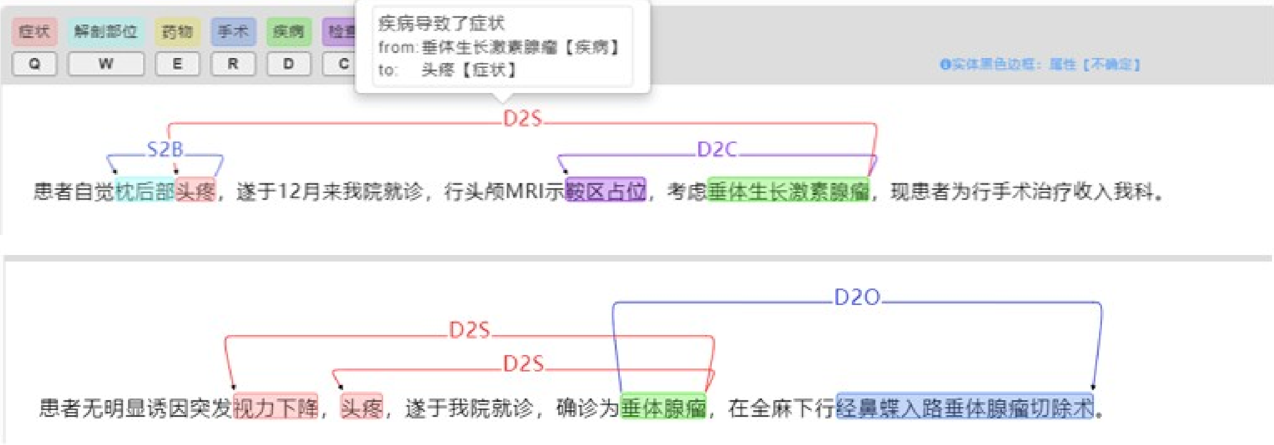
In the process of data extraction, entities and relations were first extracted from unstructured information in CEMRs. For website data, specific HTML wrappers were constructed to directly extract the triples (eg, Cushing syndrome, Symptom, Lethargy). The details are described below.

EMR Data Extraction

CEMRs include information on admission, discharge summary, disease course, and a medical record summary, among other details. Since the history of present illness (HPI) in the admission record contains a large amount of detailed patient symptoms and preliminary examination information, the HPI was selected as the main data source in our study.

The Chinese Clinical Natural Language Processing System (CCNLP) [21] developed by our team was used to annotate entities and relations in CEMRs, as shown in Figure 3. The CCNLP allows user to customize the entities and relations. According to the definition of the schema, we defined 6 types of entities and 5 types of relations in the CCNLP. Two clinicians were invited to perform annotation. The conditional random field model is embedded in the system, which can train the annotated corpus and assist in annotation. The results of the two annotators were evaluated by the consistency evaluation function of the CCNLP [22].

Figure 3. Medical text annotation using the Chinese Clinical Natural Language Processing System (CCNLP) system.



疾病导致了症状
from: 垂体生长激素腺瘤【疾病】
to: 头疼【症状】

患者自觉枕后部头疼，遂于12月来我院就诊，行头颅MRI示鞍区占位，考虑垂体生长激素腺瘤，现患者为行手术治疗收入我科。

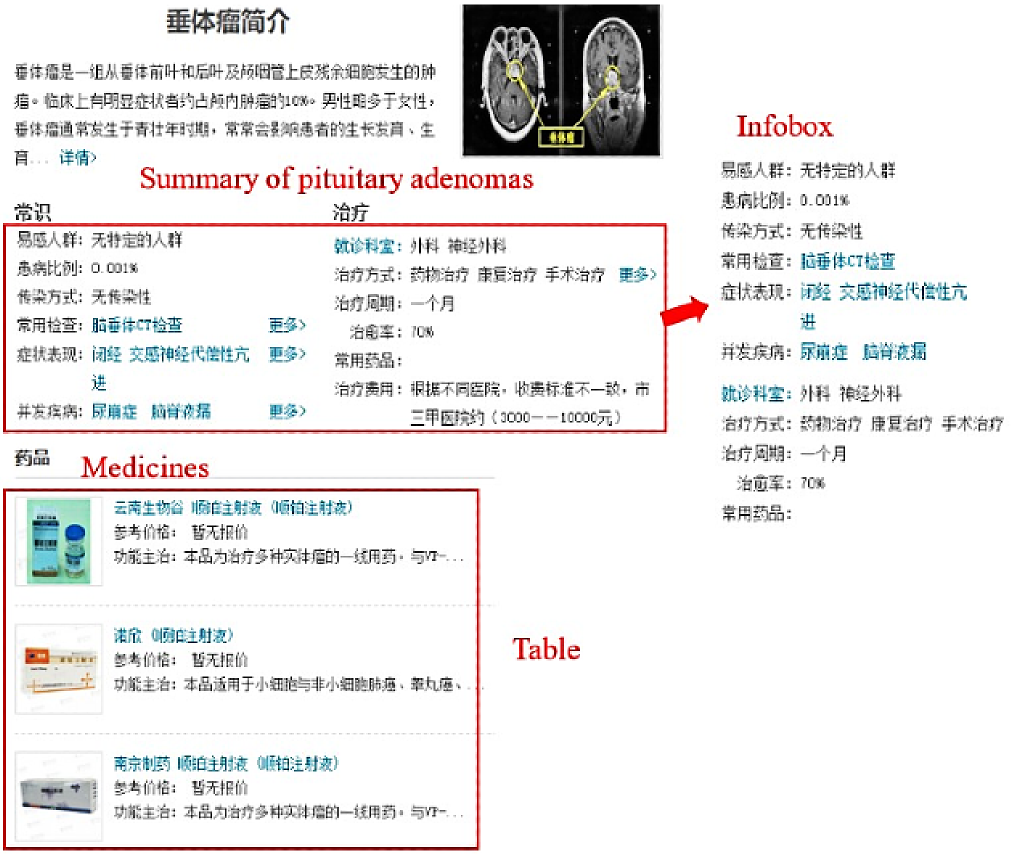
患者无明显诱因突发视力下降，头疼，遂于我院就诊，确诊为垂体腺瘤，在全麻下行经鼻蝶入路垂体腺瘤切除术。

Web Data Extraction

The web data were mainly collected from medical websites and high-quality encyclopedia websites. The extracted disease entities in the CEMRs were used as search terms on the medical websites. Since single medical website retrieval is not comprehensive, four websites with higher data quality were used: xywy [23], UpToDate [24], Baidu Encyclopedia [25], and chunyuyisheng [26]. All of these websites provide HTML pages of diseases, symptoms, treatments, and other relevant details. This enabled obtaining sufficient medical knowledge to construct the knowledge graph.

Since the websites shared similar structures, xywy was selected as an example to illustrate the details of pages and its structures used for data extraction. As shown in Figure 4, the information in "Infobox" can be directly extracted and stored as triples. The "Medicines" data in the website are stored in a tabular format. We extracted the title and first lines of the tables, which were combined as triples. Different wrappers were designed to extract information from different web pages.

Figure 4. Web page structural analysis for knowledge extraction.



垂体瘤简介

垂体瘤是一组从垂体前叶和后叶及颅咽管上皮残余细胞发生的肿瘤。临床上有明显症状者约占颅内肿瘤的10%。男性略多于女性，垂体瘤通常发生于青壮年时期，常常会影响患者的生长发育、生育… 详情>

常识

易感人群：无特定的人群
患病比例：0.001%
传染方式：无传染性
常用检查：脑垂体CT检查 更多>
症状表现：闭经 交感神经代偿性亢进 更多>
并发疾病：尿崩症 脑脊液漏 更多>

治疗

就诊科室：外科 神经外科
治疗方式：药物治疗 康复治疗 手术治疗 更多>
治疗周期：一个月
治愈率：70%
常用药品：
治疗费用：根据不同医院，收费标准不一致，市三甲医院约（3000——10000元）

药品

云南生物谷 顺铂注射液（顺铂注射液）
参考价格： 暂无报价
功能主治：本品为治疗多种实体瘤的一线用药，与VP-…

诺欣（顺铂注射液）
参考价格： 暂无报价
功能主治：本品适用于小细胞与非小细胞肺癌，睾丸癌、…

南京制药 顺铂注射液（顺铂注射液）
参考价格： 暂无报价
功能主治：本品为治疗多种实体瘤的一线用药，与VP-…

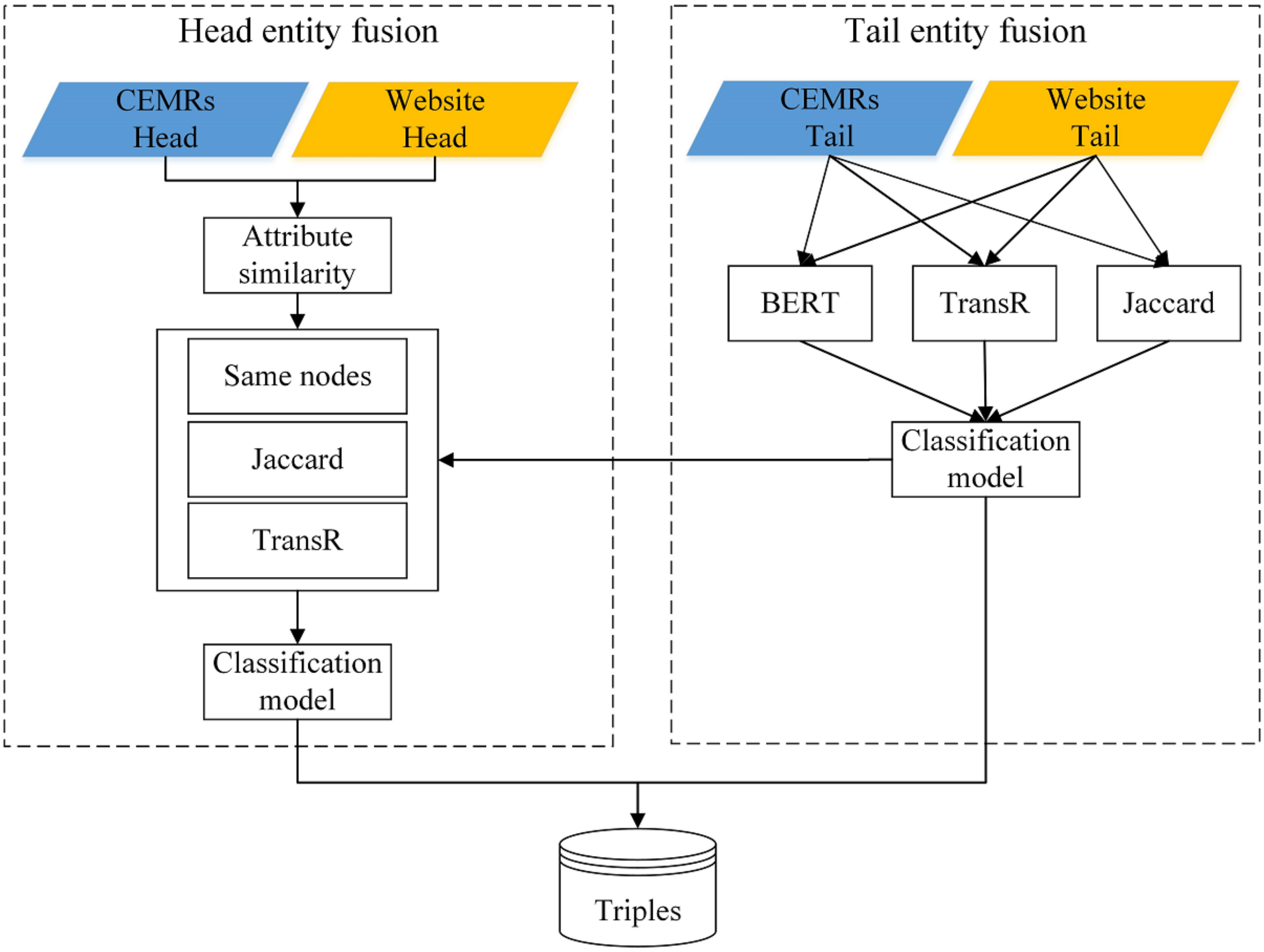
易感人群：无特定的人群
患病比例：0.001%
传染方式：无传染性
常用检查：脑垂体CT检查
症状表现：闭经 交感神经代偿性亢进
并发疾病：尿崩症 脑脊液漏
就诊科室：外科 神经外科
治疗方式：药物治疗 康复治疗 手术治疗
治疗周期：一个月
治愈率：70%
常用药品：

Data Fusion

Framework

Triples from different sources may have complements, redundancies, or even conflicts among each other. To ensure accuracy of the data in the knowledge graph, a data fusion method was proposed as shown in Figure 5. The data were fused by calculating the similarity of head entities and tail entities. The purpose of similarity calculation is to find the optimal alignment between the website entities and CEMR entities. The fusion methods were carried out in two steps. First, the similarity of tail entities (symptoms and examinations contained in both data sources) were calculated based on bidirectional encoder representations from transformer (BERT), the TransR model, and the Jaccard coefficient. Tail entity fusion enabled obtaining a more consistent entity expression. Second, the structural information of the graph was used to merge the head entities (diseases) through the TransR model, Jaccard coefficient, and the count of same nodes.

Figure 5. Data fusion framework. CEMR: Chinese electronic medical record; BERT: bidirectional encoder representations from transformer.



Tail Entity Fusion Model

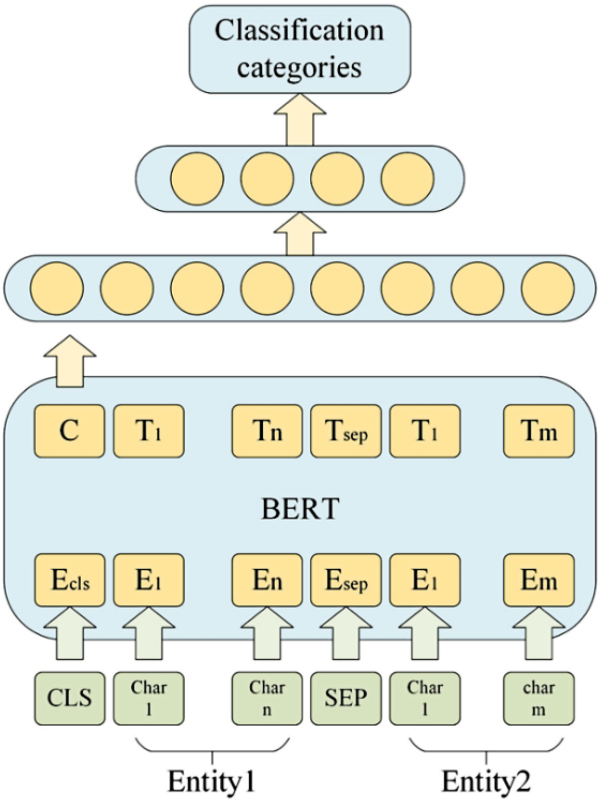
Features

In the entity fusion task, there are only two types of training results (positive and negative); therefore, this can be converted into a binary classification problem. In the tail entity fusion experiment, three different features were constructed as model inputs: semantic similarity, TransR similarity, and Jaccard similarity.

Semantic Similarity Calculation Based on BERT

A semantic model is widely used in the similarity calculation of textual data. In this study, the semantic classification model was trained with labeled data. BERT-Base, Chinese [27] was used to construct the embedding of the tail entities in CEMRs and website data, as shown in Figure 6. Tail entities can be regarded as short sentences, and the matching problem of entity pairs can be modeled as a classification task. The first output vector of the coding layer “C” is taken as the semantic representation of the entity pair. “[CLS]” represents the beginning of a sentence and “[SEP]” separates the two sentences. “E” represents the word embedding of the input character and “T” represents the contextual representation of the input character. The semantic categories are then calculated using two full connection layers: full connection layer 1 uses a tanh activation function and full connection layer 2 normalizes the probability of each class with the softmax function.

Figure 6. Semantic similarity calculation model based on bidirectional encoder representations from transformer (BERT).



Knowledge Representation Learning

Knowledge representation learning methods do not rely on textual information but rather obtain the depth characteristics of the data by mapping the entities to low-dimensional space vectors. A total of 4684 pituitary adenoma triples were used to test the data representation ability of the Trans models [28]. We evaluated the performance of the models using hits@10 (ie, the proportion of correctly aligned entities ranked in the top 10 predictions); a higher hits@10 value indicates better performance. The evaluation results were 0.27 for TransE, 0.37 for TransH, and 0.39 for TransR. Therefore, TransR was selected for knowledge representation learning. The extracted triples were used as positive examples (head [*h*], relation [*r*], tail [*t*]). For each positive triple, we randomly replaced its head entity (*h'*, *r*, *t*) or tail entity (*h*, *r*, *t'*) to generate a negative triple. A mapping matrix M_r was used to describe the relational space of relation *r*. Using the gradient descent method to update the parameters, we obtained the vector of the tail entities *trans_vec*. The cosine similarity cos was used to calculate the tail entity similarity of the two data sources, as shown in Equation 1:

$$\mathrm{Simteal\_trans}(m_i,n_i)=\mathrm{argmax}(\cos[\mathrm{trans\_vec}_{mi}],\cos[\mathrm{trans\_vec}_{ni}])$$

(1)

Jaccard Coefficient

The Jaccard coefficient was selected as the third feature of tail entity fusion. The Jaccard coefficient refers to the ratio of the number of intersection elements to the union elements in two sets; the higher the Jaccard value, the higher the similarity. We assigned each tail entity in the CEMRs and websites to sets t_1 and t_2, respectively. The Jaccard coefficient represents the ratio of the same number of Chinese characters in the two words to the total number of characters, as shown in Equation 2:

$$\mathrm{Jaccard}(t_1,t_2)=|t_1\cap t_2|/|t_1|+t_2|-|t_1\cap t_2|$$

(2)

Head Entity Fusion Model

Features

When merging head entities (diseases), the similarity of the two attributes and their structures were mainly considered. That is, if two head entities are the same, their neighboring entities should also be similar.

Attribute Similarity

Entity alignment can be performed using the alternative name attribute or the English name attribute of the disease. If the head entities in the two data sources have the same alternative name or English name, the two entities can be considered the same. For example, "垂体生长激素腺瘤" (growth hormone–secreting pituitary adenoma) has alternative names of "pituitary growth hormone secreting adenoma" and "GH adenoma." Therefore, we can align "pituitary growth hormone secreting adenoma" and "GH adenoma" to "growth hormone–secreting pituitary adenoma."

Structural Similarity Fusion Model

When the head entities cannot be aligned by the attribute, we propose using the structural similarity model to fuse entities. Three different features were chosen as the classifier model's inputs: the number of identical tail nodes, Jaccard similarity, and TransR similarity, as shown in Equation 3.

The head entity and the tail entity have a 1-N relationship. Taking two disease sets $(h_1, r_1, \sum_{i=1}^{n} t_1), (h_2, r_2, \sum_{j=1}^{n} t_2)$ from two data sources as an example, $\max \sum_{i=1}^{n} Same\_teal(t_1, t_2)$ represents the number of identical tail nodes in different sets and $Jaccard(\sum_{i=1}^{n} c_1, \sum_{j=1}^{n} c_2)$ represents the ratio of the same number of characters to the total number of characters of two sets. The order of words in the set is not considered. For the attribute similarity, the vector representation of entities was trained using the TransR model, whereas in this case, we calculated the vector of the head entity using the TransR model.

$$Simhead(h_1, h_2) = Classification(\max \sum_{i=1}^{n} Same\_teal(t_1, t_2) + Jaccard(\sum_{i=1}^{n} c_1, \sum_{j=1}^{n} c_2) + \cos \cdot TransR(h_1, h_2)) \quad \textbf{(3)}$$

After the head entities of two heterogeneous data sources were fused, the triples containing all of the disease information were obtained. Finally, to standardize the disease names in the knowledge graph, we mapped them to the ICD codes.

Results

Data Extraction

Three hundred clinical medical records and 4 portal websites were selected as data sources to construct the KGPA. Although these are all Chinese resources, our proposed approach is not dependent on a particular language and can be applied to data resources in other language in the same way. The data in CEMRs were annotated by two doctors using the CCNLP system [21]. With the consistency test function of the system, the consistency of the annotations reached 95.2%. Website data were extracted according to the wrapper defined in this study. Table 1 shows the number of all entities extracted from the two types of data sources. The concepts are abundant in websites, whereas the CEMRs included more symptom entities, which can help to expand more data types for the KGPA. The "Prefusion" column of Table 1 shows the number of all relations extracted from the two types of data sources.

Table 1. Number of relations before and after data fusion.

Relation	Head entity	Tail entity	Prefusion	After fusion
Diseases_rel_Symptom	disease	symptom	3154	1940
Diseases_rel_Surgery	disease	surgery	55	45
Diseases_rel_Medicines	disease	medicine	245	182
Diseases_rel_Examination	disease	examination	437	274
Symptoms_rel_Body structure	symptom	body	396	281
Diseases_rel_Treatment	disease	treatment	110	109
Diseases_attr_Pathogeny	disease	pathogeny	122	104
Diseases_attr_Department	disease	department	71	44
Diseases_attr_English name	disease	English name	71	42
Diseases_attr_Alternative name	disease	alternative name	23	20

Data Fusion

Two hundred medical records were randomly selected for the fusion experiment. The ratio of the training set and test set was 8:2. The experiment was trained under Windows 10, and the model based on the TensorFlow framework was used.

The proposed tail entity fusion model was used to perform entity fusion for symptoms and examinations. Before the fusion began, different entities with the same conceptual semantics extracted from different websites were merged to reduce duplication and computation. A vector representation of 768 dimensions was constructed through the Chinese BERT model, and then the similarity results were obtained by full connection layers. A 50-dimensional vector was obtained by the TransR model and the cosine similarity was used to calculate the entity pair similarity values. The Jaccard coefficient was used as a numerical feature. These three results were taken as features into the classification model. Three different classification models were adopted for training: logistic regression, decision tree, and neural network. The results are shown in Table 2. The neural network showed the best performance.

Subsequently, the triples completed by the tail entity fusion model were used for the head entity fusion experiment. A total of 65 head entities were fused between CEMRs and websites. Among them, 17 entities could be directly mapped by disease name, 6 entities could be fused by attribute (eg, growth hormone–secreting pituitary adenoma, pituitary microadenoma, Cushing syndrome, hypothyroidism), and 42 head entities were fused based on the proposed structural similarity fusion model. The three classification models above were used for training. As shown in Table 2, the decision trees performed better when fusing head entities because the data inputs to the model were smaller than the fusing tail entities. With the increase of data volume, the advantages of the neural network were reflected in the fusion of tail entities.

Additionally, we divided the features into four variants for an ablation study. We selected logistic regression as the classification model to explore the contribution of different features to the model, and these results are also shown in Table 2. These three features had nearly the same contributions to the model in the head entity fusion. For a specific disease knowledge graph, the Jaccard similarity feature played a major role in the tail entity ablation experiment, and the features based on BERT and TransR simply contributed by fine-tuning the model.

Table 3 shows that our proposed model has higher accuracy than previous models. Compared with previous models, we divided the entities into head entities and tail entities and fused them according to different characteristics. Different concepts were considered separately in the step-by-step fusion process, which improved the precision of the fusion.

Table 2. Head and tail fusion model performance.

Fusion model	Precision (%)	Recall (%)	F-score (%)
Head entity fusion			
Linear regression models			
Ja[a]+TransR	83.37	84.06	83.71
Sa[b]+TransR	83.37	84.55	83.95
Ja+Sa	83.85	84.55	84.19
Ja+Sa+TransR	83.92	84.61	84.26
Neural network	97.29	97.03	97.16
Decision tree	97.47	97.18	97.32
Tail entity fusion			
Linear regression models			
BERT[c]+TransR	61.73	61.74	61.73
Ja+BERT	95.76	95.83	95.79
Ja+TransR	95.89	95.93	95.90
Ja+BERT+TransR	95.92	95.94	95.93
Neural network	98.43	98.72	98.57
Decision tree	98.18	98.05	98.11

[a] Ja : Jaccard similarity.

[b] Sa: identical tail nodes in different sets: $\max\sum_{i=1}^{n} Same\_teal(t_1,t_2)$.

[c] BERT: bidirectional encoder representations from transformer.

Table 3. Model comparison.

Model	Research field	Method	F1-score
Ruan et al [29]	Symptom	Align entities according to the string similarities of the entity names and attribute values	—[a]
Yang et al [30]	Disease, medicine	Align entities according to the entity's attribute types ($attr_{bool}$, $attr_{numeric}$, $attr_{string}$, $attr_{time}$)	0.60
Sun et al [31]	Disease, medicine, symptom	Character similarity of entity pairs and degree centrality of entities in the graph	0.76
Liu et al [32]	Disease, medicine, examination	Semantic classification model based on pretrained BERT[b]	0.83
Our model	Symptom, examination, disease	Multifeature learning based on head-and-tail entities	0.97

[a] Not provided.

[b] BERT: bidirectional encoder representations from transformer.

The triples obtained after data fusion were stored and visualized in Neo4j [33]. The KGPA contained 1789 entities and 3041 pairs of relations of 73 pituitary adenoma–related diseases. For a knowledge graph, accuracy is of great importance. However, there is currently no gold standard for pituitary adenoma knowledge graph validation. To evaluate the quality of the knowledge graph, the accuracy of triples was used as an indicator. Three hundred triples were randomly sampled and each triple was manually evaluated by two physicians; the accuracy reached 95.4%.

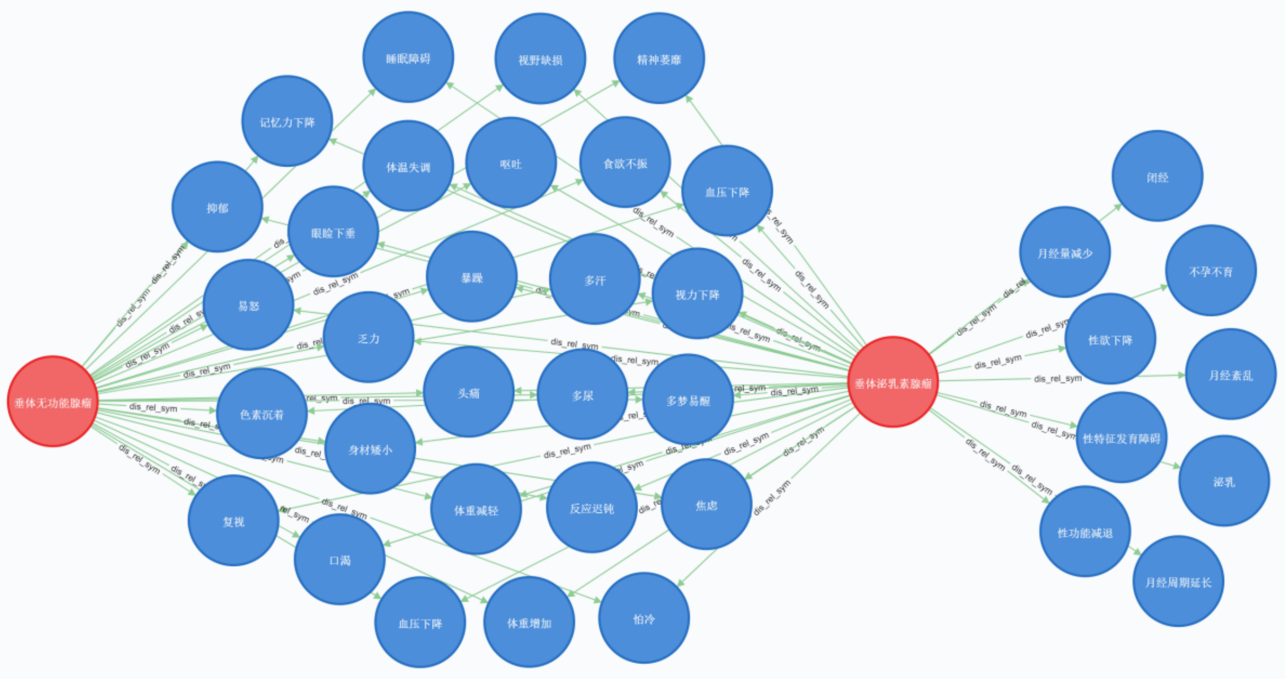
Discussion

Principal Findings

A knowledge graph was constructed by mining CEMRs and web resources. In the process of KGPA construction, to solve the problem of knowledge duplication between heterogeneous data sources, we proposed a head-and-tail entity fusion model. The model showed good performance on the fusion of medical data.

The KGPA was proven to be effective when displaying the typical symptoms of pituitary adenoma–related diseases. For example, the query for symptoms of disease “prolactin (PRL)-secreting pituitary adenomas” differed from the query for the disease “nonfunctioning pituitary adenoma” using the following query in Cypher: “MATCH (p:dis{disease: 垂体泌乳素腺瘤})-[:dis_rel_sym]->(n), (m)<-[:dis_rel_sym] -(q:dis{disease:垂体无功能腺瘤}), WHERE (m)<>(n), RETURN p,n,q.” As shown in Figure 7, the entities in the middle of the graph are symptoms of both diseases and the entities on the right are typical symptoms unique to the disease “PRL-secreting pituitary adenomas.”

Figure 7. Differences of typical symptoms between “prolactin-secreting pituitary adenomas” and “nonfunctioning pituitary adenoma” in the knowledge graph for pituitary adenoma.



Searching for the KGPA by Cypher, we found that most pituitary adenoma–related diseases have the following basic symptoms: headache, vision problems, fatigue, slow reaction, mood problems, changes in height and weight, changes in appetite, and changes in sleep. Nonfunctioning pituitary adenoma has all of these basic symptoms listed above. In addition to the basic symptoms, pituitary thyroid-stimulating hormone adenoma is also associated with symptoms of goiter, palpitation, and exophthalmos. The typical symptoms of PRL-secreting pituitary adenomas are associated with the reproductive system, decreased libido, and menstrual changes in women. The typical symptoms of pituitary growth hormone adenoma are altered facial features, enlarged hands and feet, snoring, and metabolic disorders. Cushing syndrome is characterized by obesity, altered skin color, increased hair, and edema. Based on clinicians’ feedback on the use of the KGPA, the knowledge in the KGPA was consistent with the actual clinical situation. The KGPA will be useful for clinical interns in diagnosis and treatment, and may also be helpful for medical students to quickly master knowledge of pituitary adenoma–related diseases.

Limitations

The KGPA was constructed by integrating CEMRs and web data related to pituitary adenoma. However, since we only focused on pituitary tumors, the data volume was relatively small. In the next step, we plan to try to extend the method proposed in this study to the entire neurosurgery field or even larger fields and apply the knowledge graph to clinical practice.

Conclusion

This study shows that entities and relations extracted from heterogeneous data sources such as CEMRs and health websites can be used to construct a knowledge graph after entity fusion. The head-and-tail entity fusion model proposed in this paper achieved 97% in accuracy, which is higher than that reported for previous models. The KGPA constructed in this study can be used to discover the knowledge hidden in the source text, such as typical symptoms unique to the disease “PRL-secreting pituitary adenomas.” Based on clinicians’ feedback, the knowledge in the KGPA was consistent with the actual clinical situation. The knowledge graph constructed will be useful and helpful for patients, medical students, and interns to assist in obtaining information for symptoms, diagnosis, treatment, and disease pathogenesis.

Acknowledgments

This research has been funded by the Science and Technology Innovation 2030-Major Project (2020AAA0104902), the Chinese Academy of Medical Sciences Initiative for Innovative Medicine (2017-I2M-3-014), the Chinese Academy of Medical Sciences and Peking Union Medical College Fundamental Scientific Research Funds Project of the Central Public Welfare Research Institution (2018PT33005), and the Hunan Provincial Key Research and Development Program (2020SK2089).

Authors' Contributions

AF designed the methods, analyzed the results of experiments, and drafted the paper. PL, JH, and WZ extracted the data and performed the data fusion. MF collected the electronic medical records and annotated the dataset. MF and HR evaluated the pituitary adenoma knowledge graph. XC supervised the research and revised the paper. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Osamura R, Egashira N, Miyai S, Yamazaki M, Takekoshi S, Sanno N, et al. Molecular pathology of the pituitary. Development and functional differentiation of pituitary adenomas. Front Horm Res 2004;32:20-33. [doi: 10.1159/000079036] [Medline: 15281338]
2. Kinoshita Y, Tominaga A, Usui S, Arita K, Sugiyama K, Kurisu K. Impact of subclinical haemorrhage on the pituitary gland in patients with pituitary adenomas. Clin Endocrinol (Oxf) 2014 May;80(5):720-725. [doi: 10.1111/cen.12349] [Medline: 24125536]
3. Kim JP, Park BJ, Kim SB, Lim YJ. Pituitary apoplexy due to pituitary adenoma infarction. J Korean Neurosurg Soc 2008 May;43(5):246-249 [FREE Full text] [doi: 10.3340/jkns.2008.43.5.246] [Medline: 19096606]
4. Kaushik C, Ramakrishnaiah R, Angtuaco EJ. Ectopic pituitary adenoma in persistent craniopharyngeal canal. J Comput Assist Tomogr 2010;34(4):612-614. [doi: 10.1097/rct.0b013e3181dbe5d1]
5. Byambasuren O, Yang Y, Sui Z, Dai D, Chang B, Li S, et al. Preliminary study on the construction of Chinese medical knowledge graph. J Chinese Inf Process 2019;33(10):1-9 [FREE Full text]
6. Li L, Wang P, Yan J, Wang Y, Li S, Jiang J, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med 2020 Mar;103:101817. [doi: 10.1016/j.artmed.2020.101817] [Medline: 32143785]
7. Gong F, Chen Y, Wang H, Lu H. On building a diabetes centric knowledge base via mining the web. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):49 [FREE Full text] [doi: 10.1186/s12911-019-0771-6] [Medline: 30961582]
8. Ernst P, Siu A, Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 2015 May 14;16:157 [FREE Full text] [doi: 10.1186/s12859-015-0549-5] [Medline: 25971816]
9. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. Sci Rep 2017 Jul 20;7(1):5994. [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
10. Zhang Q, Lian B, Cao P, Sang Y, Huang W, Qi L. Multi-source medical data integration and mining for healthcare services. IEEE Access 2020;8:165010-165017. [doi: 10.1109/access.2020.3023332]
11. Wang C, Wang H, Zhuang H, Li W, Han S, Zhang H, et al. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree. J Biomed Inform 2020 Nov;111:103583. [doi: 10.1016/j.jbi.2020.103583] [Medline: 33010427]
12. Li W, Zhang X, Wang Y, Yan Z, Peng R. Graph2Seq: fusion embedding learning for knowledge graph completion. IEEE Access 2019;7:157960-157971. [doi: 10.1109/access.2019.2950230]
13. Li N, Yang Z, Luo L, Wang L, Zhang Y, Lin H, et al. KGHC: a knowledge graph for hepatocellular carcinoma. BMC Med Inform Decis Mak 2020 Jul 09;20(Suppl 3):135 [FREE Full text] [doi: 10.1186/s12911-020-1112-5] [Medline: 32646496]
14. Nickel M, Murphy K, Tresp V, Gabrilovich E. A review of relational machine learning for knowledge graphs. Proc IEEE 2016 Jan;104(1):11-33. [doi: 10.1109/jproc.2015.2483592]
15. Unified Medical Language System (UMLS). National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/index.html [accessed 2021-02-09]
16. SNOMED International. URL: https://www.snomed.org/ [accessed 2021-02-09]
17. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011 Sep 01;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
18. Yang J, Guan Y, He B, Qu CY, Yu QB, Liu YX, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records. J Softw 2016:2725-2746. [doi: 10.13328/j.cnki.jos.004880]
19. Yang C, Hu J, Fang A. Study on the building of clinical text natural language processing system—taking cTAKES as an example. J Med Inform 2018;39(12):48-53.

20. Su X, Fan K, Shi W. Privacy-preserving distributed data fusion based on attribute protection. IEEE Trans Ind Inf 2019 Oct;15(10):5765-5777. [doi: 10.1109/tii.2019.2912175]
21. Chinese Clinical Natural Language Processing System (CCNLP). 2021. URL: http://ccnlp.imicams.ac.cn/ [accessed 2021-02-09]
22. Hu J, Fang A, Zhao W, Yang C, Ren H. Annotating Chinese e-medical record for knowledge discovery. Data Anal Knowl Discov 2019;3(7):123-132.
23. xywy. URL: http://www.xywy.com/ [accessed 2020-12-20]
24. UpToDate. URL: https://www.uptodate.cn/home/ [accessed 2020-12-20]
25. Baidu Encyclopedia. URL: https://baike.baidu.com/ [accessed 2020-12-20]
26. chunyuyisheng. URL: https://www.chunyuyisheng.com/ [accessed 2020-12-20]
27. Devlin J, Chang MW, Lee K. BERT: Pre-train-ing of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2021-07-07]
28. Lin H, Liu Y, Wang W, Yue Y, Lin Z. Learning entity and relation embeddings for knowledge resolution. Procedia Comput Sci 2017;108:345-354. [doi: 10.1016/j.procs.2017.05.045]
29. Ruan T, Wang M, Sun J, Wang T, Zeng L, Yin Y, et al. An automatic approach for constructing a knowledge base of symptoms in Chinese. J Biomed Semantics 2017 Sep 20;8(Suppl 1):33 [FREE Full text] [doi: 10.1186/s13326-017-0145-x] [Medline: 29297414]
30. Fu Y, Liu M, Qiao R. Construction of Chinese knowledge graph of heart disease. J Wuhan Univ 2020;66(3):261-267. [doi: 10.14188/j.1671-8836.2018.0217]
31. Sun Q. Knowledge extraction and alignment for respiratory disease. Harbin Institute of Technology. 2019. URL: https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD 202001&filename=1019646460.nh/ [accessed 2021-03-30]
32. Liu X, Jin J, Ruan T, Gao D, Yin Y, Ge X. Construction of an open dataset for clinical event graph. J Chinese Inf Process 2020;11:37-48.
33. Balaur I, Mazein A, Saqi M, Lysenko A, Rawlings CJ, Auffray C. Recon2Neo4j: applying graph database technologies for managing comprehensive genome-scale networks. Bioinformatics 2017 Apr 01;33(7):1096-1098 [FREE Full text] [doi: 10.1093/bioinformatics/btw731] [Medline: 27993779]

Abbreviations

BERT: bidirectional encoder representations from transformer
CCNLP: Chinese Clinical Natural Language Processing System.
CEMR: Chinese electronic medical record
EMR: electronic medical record
HPI: history of present illness
ICD: International Classification of Diseases
KGPA: knowledge graph for pituitary adenoma
PRL: prolactin

Edited by T Hao; submitted 25.02.21; peer-reviewed by S Zhu, H Lin; comments to author 12.03.21; revised version received 11.04.21; accepted 30.05.21; published 22.07.21

Please cite as:
Fang A, Lou P, Hu J, Zhao W, Feng M, Ren H, Chen X
Head and Tail Entity Fusion Model in Medical Knowledge Graph Construction: Case Study for Pituitary Adenoma
JMIR Med Inform 2021;9(7):e28218
URL: https://medinform.jmir.org/2021/7/e28218
doi: 10.2196/28218
PMID: 34057414