Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(4): 6776–6799.
DOI: 10.3934/mbe.2023292
Received: 02 September 2022
Revised: 14 January 2023
Accepted: 17 January 2023
Published: 06 February 2023

Research article

Construction and application of Chinese breast cancer knowledge graph based on multi-source heterogeneous data

Bo An[1,2,*]

[1] Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing 100732, China

[2] Beijing Academy of Artificial Intelligence, Beijing 100084, China

* **Correspondence:** Email: anbo@cass.org.cn.

Abstract: The knowledge graph is a critical resource for medical intelligence. The general medical knowledge graph tries to include all diseases and contains much medical knowledge. However, it is challenging to review all the triples manually. Therefore the quality of the knowledge graph can not support intelligence medical applications. Breast cancer is one of the highest incidences of cancer at present. It is urgent to improve the efficiency of breast cancer diagnosis and treatment through artificial intelligence technology and improve the postoperative health status of breast cancer patients. This paper proposes a framework to construct a breast cancer knowledge graph from heterogeneous data resources in response to this demand. Specifically, this paper extracts knowledge triple from clinical guidelines, medical encyclopedias and electronic medical records. Furthermore, the triples from different data resources are fused to build a breast cancer knowledge graph (BCKG). Experimental results demonstrate that BCKG can support knowledge-based question answering, breast cancer postoperative follow-up and healthcare, and improve the quality and efficiency of breast cancer diagnosis, treatment and management.

Keywords: knowledge graph; medical knowledge graph; information etraction; deep learning; pre-trained language model

1. Introduction

The medical knowledge graph is a formal and semantic description that reveals the relationship among medical entities such as disease, symptom, medicine, and surgery. Building high-quality medical knowledge graphs can effectively improve the management and utilization of medical resources, and greatly promote the development of intelligent medical applications such as knowledge-based question answering, intelligent search, and decision support [1]. The disease knowledge graph is an essential

part of the medical knowledge graph [2], which can realize an accurate and comprehensive collection of specific disease knowledge and can support specialized medical intelligence applications, such as clinical-assisted decision-making, drug development, knowledge-based question answering, intelligent marketing, intelligent underwriting claims, etc. From the research status of the specialized disease knowledge graph, there are specialized disease knowledge graphs in English, taking breast cancer knowledge base * as an example, which includes knowledge and medical record data related to breast cancer disease. This knowledge plays a crucial role in the diagnosis and treatment, health management and clinical research of breast cancer [3]. Most Chinese medical knowledge graph researches focus on general knowledge graph [4], and there is a lack of disease-specific knowledge graph research and knowledge graph construction. The general medical knowledge graph is oriented to all diseases. It contains a huge amount of triples which is challenging to review manually and has limitations in terms of accuracy and coverage for specific diseases, etc., which to a certain extent limits the application of artificial intelligence in the field of specialized diseases [1].

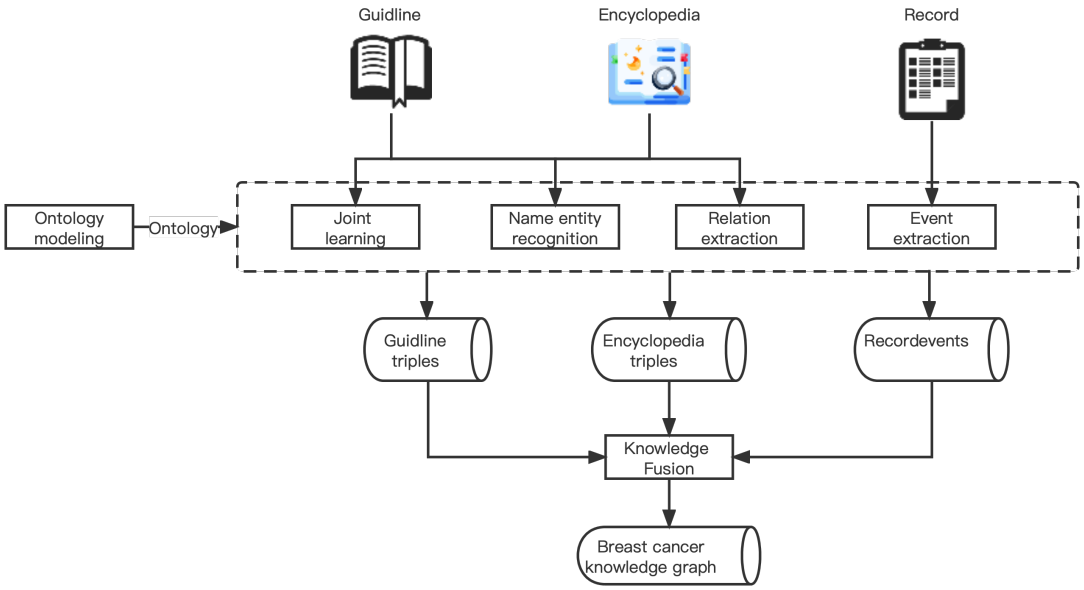

Figure 1. The illustration of the proposed framework.

To solve the issues, this paper takes breast cancer as an example and proposes a method to build a disease knowledge graph that integrates multi-source heterogeneous data. This method can apply various knowledge extraction methods, extract breast cancer-related knowledge from multi-source heterogeneous data, and build a knowledge graph of breast cancer through knowledge fusion. The research framework is shown in Figure 1. Firstly, according to the characteristics of data from different sources, this paper designs and implements the methods of named entity recognition and relation extraction based on joint learning [5–7], table extraction [8]. Event extraction [9,10] to extract knowledge from breast cancer clinical guidelines text, medical encyclopedia, and medical records [11]. Then, the knowledge fusion method is utilized to build a knowledge graph of specific diseases [12]. In this paper, we employ accuracy, recall and F1-value to evaluate the quality of the knowledge graph. The main

*https://wasp.cs.vu.nl/BreastCancerKG/

contributions of this paper can be summarized as follows:

- A method is proposed to construct the breast cancer knowledge graph based on multi-source heterogeneous data, which can extract entity, relation, attribute, and event information from unstructured text, semi-structured text and table, such as clinical guidelines, medical encyclopedias, and medical record data.
- Typical data sources in the medical field, such as clinical guidelines, medical encyclopedias and electronic medical records, are introduced in the knowledge graph construction. The article optimizes the traditional knowledge extraction and fusion knowledge graph algorithms, and improves them in terms of accuracy and effectiveness, which have high practical value.
- Combining with specific application scenarios in the medical field, the constructed knowledge graph is applied to medical question answering, and medical record retrieval.

2. Related work

2.1. The concept and evolution of knowledge graphs

The knowledge graph is a structured form to describe concepts, entities, and their relation in the objective world, which facilitates the understanding of humans and machines and reduces the difficulty of knowledge application [13, 14]. Google first proposed knowledge graph in 2012, knowledge graph has gained wide attention in industry and academia and has played an essential role in information retrieval, question answering, semantic understanding, and intelligent healthcare [1, 15, 16]. Generally speaking, knowledge graphs are usually divided into generic knowledge graphs and domain knowledge graphs, and generic knowledge graphs cover knowledge in various domains, which are usually large and are generated by automatic extraction or co-editing methods. For example, Carnegie Mellon University developed NELL [17], a typical representative of general-purpose knowledge graphs. NELL uses automatic extraction to build knowledge graphs, constantly crawling text from the Internet and extracting knowledge from the text by named entity recognition and relation extraction to form knowledge graphs. Then, for example, Wikidata [18, 19] is a large-scale general knowledge graph formed by collating data based on Wikipedia and other data, supporting crowdsourced collaborative editing, and can support more than 350 languages involving 25 million entities. It can be seen that the general knowledge graph has the characteristics of large-scale and broad knowledge coverage. However, the shallow specialization and the lack of manual verification lead to the shortage of the general knowledge graph in accuracy and coverage for domain-specific problems.

More and more scholars focus on domain knowledge graphs to improve the domain depth of knowledge graphs. Compared with general-purpose knowledge graphs, the scope of domain knowledge graphs knowledge is oriented to a specific professional domain, such as financial knowledge graphs [20], legal knowledge graphs [21], poetry knowledge graphs [22], and medical knowledge graphs [23]. Domain knowledge graphs usually only cover knowledge in the domain. The scale of knowledge is smaller than that of general-purpose knowledge graphs. However, the coverage of knowledge graphs in the domain will be more complete, such as the OpenKG [24], which has built and open-sourced several domain knowledge graphs involving more than 15 domain categories such as medical, travel, financial, legal, urban, and travel, etc., and is oriented to specific professional domain knowledge graphs, smaller in scale but with higher degree of specialization and domain specificity.

2.2. Medical knowledge graph and applications

Medical knowledge graphs are semantic descriptions that reveal the relations between medical concepts such as disease, symptoms, drugs, surgery, protection, and rehabilitation. The medical knowledge graph is a fundamental and critical work for question answering, retrieval systems, intelligent medical decision-making, patient portraits, and medical record integration, and it has received extensive attention. [25] constructed a COVID-19 knowledge graph based on the public datasets to obtain new crown epidemic and epidemic data, and equipped with a question answering system. Shanshan Zhai [25] combined the knowledge graph with faceted retrieval and constructed an online faceted retrieval model for chronic diseases based on the medical knowledge graph. Several medical-related knowledge graphs have been released, such as the large-scale medical knowledge graph extracted from PubMed literature by IEDA [†], the medical knowledge graph extracted based on clinical data by Omaha mapping [‡], and Pengcheng laboratory released a Chinese medical knowledge graph Cmekg [4].

Knowledge extraction is the core part of building knowledge graphs, which involves techniques such as named entity recognition and relation extraction [26, 27]. The main knowledge extraction methods are rule-based extraction methods, statistical learning-based extraction methods, and deep learning-based extraction methods [28]. With the development and broad application of deep learning and large-scale pre-trained language models, pre-trained language models are widely used for tasks such as entity recognition and relation extraction [29, 30]. The traditional approach models named entity recognition and relation extraction as a pipeline task, i.e., recognizing entities in text first and performing relation extraction based on the obtained entity pairs, which suffers from error propagation [6]. To address these shortcomings, researchers proposed a joint learning-based named entity recognition and relation extraction approach [6, 31], which models named entity recognition and relation extraction in the same task and model. The model utilizes entity information to identify relations while also using relation information to constrain the results of named entity recognition to obtain better knowledge extraction results.

The above analysis shows that with the penetration of knowledge graphs into the medical field, medical knowledge graphs have made significant progress from both research and application aspects. However, constrained by the quality and scale of medical knowledge graphs, some things could be improved in applications [32]. There are two main reasons: on the one hand, there are limitations in the knowledge covered: the medical knowledge graph currently tries to cover all diseases, and due to the complexity of the medical field, the knowledge covered by the existing medical knowledge graph is still very limited [33]. For example, the diseases related to breast cancer in Cmekg include only a few diseases, such as 'breast carcinoma in situ', but there are more than 700 diseases belonging to the category of breast cancer, such as 'invasive breast cancer'. On the other hand, the data sources are single: the current knowledge graphs are mainly extracted from a single data resource, but in the medical field, there are multiple high-quality knowledge sources, such as disease guidelines (unstructured text), medical encyclopedias (semi-structured data), medical record data (semi-structured data), etc., which can complement each other to build a complete disease knowledge graph. This paper takes breast cancer as an example. This paper proposes a method for constructing a breast cancer knowledge graph by fusing heterogeneous data from multiple sources, which can compose several different knowledge extraction methods, extract breast cancer-related knowledge from heterogeneous data, and build a breast cancer-

[†]https://idea.edu.cn/news/20220128095607.html
[‡]http://www.omaha.org.cn/index.php?m=hita

specific knowledge graph by knowledge fusion technology, which can alleviate the above problems to a certain extent and improve the applicability and universality of the knowledge graph.

3. Knowledge graph construction methods

In this paper, we take clinical guidelines (diagnosis and treatment, follow-up, etc.), medical encyclopedias (You Lai Encyclopedia [§], Medical Encyclopedia [¶]) and medical records as the data resources. And we extract knowledge from these resources to build the breast cancer knowledge graph. The algorithm mainly consists of 3 parts, namely: 1) ontology modelling, 2) knowledge extraction, and 3) knowledge fusion and construction of breast cancer knowledge graphs. This section introduces the construction methods.

3.1. Ontology modeling

Knowledge ontology is a formal representation of a concept, an abstract model of knowledge of the world, abstracted into a machine-understandable and processable form. Ontology modelling usually consists of two approaches: top-down and bottom-up. Top-down ontology modelling is mainly done collaboratively by domain experts and knowledge graph experts and is usually oriented to specific domains and focuses on expert knowledge. Bottom-up ontology modelling is usually done automatically or semi-automatically based on data and focuses more on abstraction from the knowledge data itself. The ontology is the primary data model of the knowledge graph, specifying the entity types, relation types, and attribute types involved in the knowledge graph, and is usually jointly determined by domain experts and knowledge graph experts. In this paper, we invite breast cancer experts from the Cancer Hospital of China Medical University to participate in the development of breast cancer ontology structure, including 'clinical observation', 'surgical operation', 'examination', 'specimen', etc. The structure of breast cancer ontology includes 28 categories of entities, including 'clinical observation', 'surgical operation', 'examination' and 'specimen'. For example, the entity 'Examination' also includes 'imaging', 'Laboratory', 'Pathology' and 'Endoscopy', 'endoscopy' and other six types of entities. Entities are associated with each other through 'relations', for example, between 'disease' and 'imaging' through 'imaging'. The relationship between entities is established by relations, such as between 'disease' and 'imaging test' by 'imaging test', which forms (disease, imaging test, imaging test). In addition to the relationship, the entity also contains some attribute information, such as whether the disease is 'hereditary', 'common', etc. Entities are related to other entities through relationships, and entities are related to specific attribute values through attributes, usually numbers, text, boolean values, etc. In this paper, we use Protege [34] as an ontology modeling tool, and the ontology structure formed is shown in Figure 2. Where the square's represent entities, the square's represented by '+' sign by subclass entities, and the connecting lines represent the relationship between entities.

To further enhance the application value and potential of the knowledge graph of breast cancer, the article adds the medical record data to the knowledge graph. It adds new entities such as 'patient', 'hospital' and 'doctor'. In this article, we added the patient, hospital, and physician entities into the knowledge graph, and the patient is associated with the disease and the clinical observation. At the same time, the patient's medical history data are extracted as events. Specifically, the article selects the

[§]https://www.youlai.cn/
[¶]https://www.yixue.com/

information that has a significant impact on breast cancer diagnosis and treatment for event modelling, such as 'surgery event', 'consultation event', 'hospitalization event', 'CT examination event', and 'CT examination event'. The article selected 15 types of events, such as 'surgery event', 'consultation event', 'hospitalization event', 'CT examination event', and 'first disease course event'. The event usually contains event type and event element'. By extracting, displaying, and applying the patient's medical record information in an event graph, the combination of patient cases and knowledge graph is conducive to applying knowledge graph in clinical decision support, drug development, and medical record retrieval. Medical experts jointly determine the schema of events to identify events' types and constituent elements.

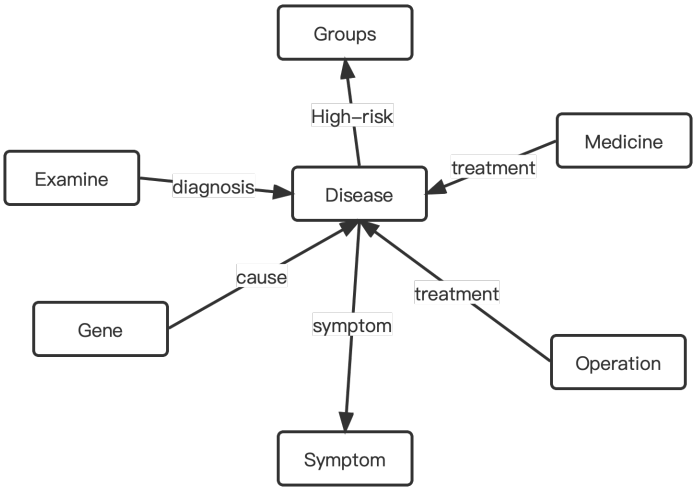

Figure 2. The illustrate of ontology of BCKG.

3.2. Knowledge extraction

After the ontology structure is modelled, knowledge extraction from data sources needs to be carried out to extract knowledge in the form of triple, which contains three parts: head entities, relations and tail entities, such as (breast cancer, symptoms, breast pain). The quality of data sources has an essential impact on the final extracted knowledge graph quality. In this paper, we expect to construct high-quality breast cancer knowledge graphs and apply them in professional applications such as clinical aid diagnosis, medical record retrieval, and postoperative follow-up. Therefore, this paper extracts knowledge from three kinds of high-quality data resource: 1) clinical guidelines (diagnosis, follow-up, etc.): including unstructured text data, and semi-structured table data; 2) medical encyclopedias (Youlai Medical Encyclopedia): containing structured Infobox and unstructured text; 3) clinical electronic medical records: unstructured text data. The three data sources contain unstructured, semi-structured, and structured heterogeneous data. Due to the different formats and characteristics of these three data, different methods will be utilized for knowledge extraction in this paper, which will be introduced separately in this section.

This paper employs different extraction methods according to the characteristic of data resources. And we will detail describe these methods in the following section.

3.2.1. Knowledge extraction from clinical guidelines

Clinical guidelines are the most crucial instruction for disease diagnosis, treatment, postoperative, follow-up, etc. [35]. They are an essential basis for disease diagnosis and treatment and a core data source for building the medical knowledge graph. Clinical guidelines usually contain different directions (diagnosis and treatment, follow-up, etc.) and different publishing organizations (Health and Welfare Commission, Chinese Society of Clinical Oncology, etc.). Guidelines often contain data with multiple structures, and Figure 3 gives three different types of data (text, table, and combined graphic and text) contained in the guidelines. The article combines the structural characteristics of the different data types and employs different knowledge extraction methods.

Figure 3. The illustration of clinical guidelines.

3.2.2. Joint learning model for knowledge extraction of guide texts

The guideline text contains the entities and the relations, as shown in Figure 3(a). An example of knowledge extraction based on joint learning is given in Figure 4. The model outputs both entity and relation information. As shown in Figure 4, the model identifies the entities 'breast cancer' and 'dimple disease' and determines the relationship between them based on the sentence's content. The model identifies the entities 'breast cancer' and 'dimple disease' and determines the relationship between them based on the sentence content. At the algorithm level, the two main tasks are named entity recognition and relation extraction. The traditional approach employs a pipeline approach to model the two tasks separately, i.e., a model is utilized to identify named entities first. Then another model is employed to determine whether there is a relation between entity pairs to obtaining knowledge triple. The pipeline-based knowledge extraction approach is simple to implement, but suffers from problems such as error transmission and model information not being shared.

To address this problem, the paper designs and implements a joint learning-based knowledge extraction method for guideline text, which models entity recognition and relation extraction as a Token Pair Classification (TPC) task, and the TPC algorithm is elaborated below with a specific example. m relations are modeled by TPC using m $n \times n$ matrices M, i.e., relation i is represented using matrix M_i, where n is the number of characters in the text. Figure 4 shows the modelling method of TPC for knowledge extraction of the text 'Taxol can treat breast cancer'. As shown in the figure, the text contains 10 characters, so each relation corresponds to a 10×10 matrix. The horizontal and vertical

coordinates of each element of the matrix correspond to the character's position in the sentence. Each element in the matrix is the algorithm's classification target, and four classification labels correspond to the following meanings. 1) '-' indicates no starting information about the entity. 2) 'HB' indicates the start character of the head entity. 3) 'HE indicates the header entity's end character and the tail entity's start character. 4) 'TE' indicates the end character of the tail entity. In Figure 4, 'HB' and 'HE' corresponds to the string 'Taxol' for the head entity, 'HE' and 'TE' for the tail entity. and 'TE' corresponds to 'breast cancer' as the tail entity, and the current matrix corresponds to 'treatment'. Based on this character pair classification result, we can obtain the triple (taxol, treatment, breast cancer). The method models entity recognition and relation extraction as a unified task utilizing character pair classification and recognizes both entities and inter-entity relationships in the text. Since the model employs a unified parameter update strategy, the results of relation extraction can be fed directly to the model for updating the overall parameters and thus optimizing entity recognition results.

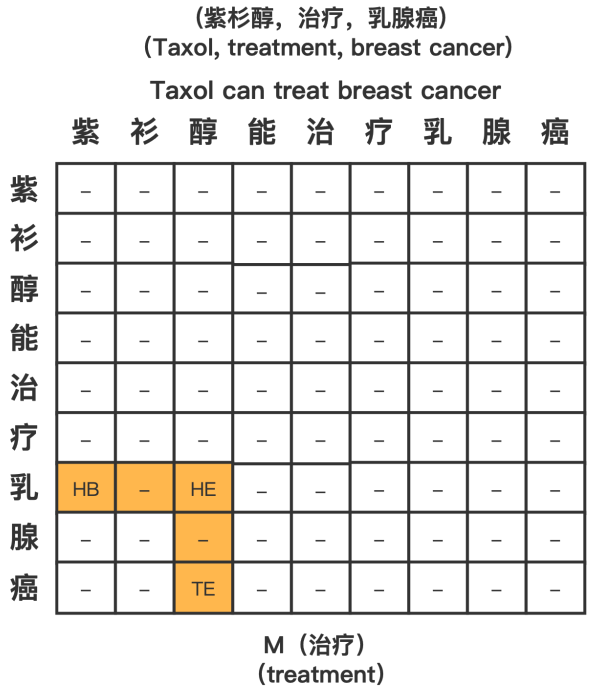

Figure 4. The illustration of joint learning.

3.2.3. Table knowledge extraction

Clinical guidelines often contain many tables, which distill medical experts' knowledge of disease diagnosis and treatment. For example, the tests required to confirm the diagnosis of early breast cancer are shown in Table 1. The data shows that the table headers in the guideline tables contain important entity and relation information, such as the disease entity 'early breast cancer' and the relationship 'confirmatory tests' in Table 1. The content of the table usually lists the entities that correspond to the entities in the header, such as the examination entity 'bilateral mammograms', forming a triple of knowledge (early breast cancer, confirmatory examination, bilateral mammograms).

Table 1. Diagnostic examination of early breast cancer.

Position	Diagnostic
原发肿瘤评估 (Primary tumor assessment)	1）体格检查 (physical examination) 2）双侧乳腺 X 线摄片 (Bilateral mammography) 3）超声 (Breast ultrasonic) 4）乳腺磁共振 (Breast MRI) 5）空芯针穿刺 (Hollow core needle puncture)
区域淋巴结评估 (Regional lymph node assessment)	1）体格检查 (physical examination) 2）双侧乳腺 X 线摄片 (Bilateral mammography) 3）可疑病灶空芯针穿刺 (Hollow core needle puncture for suspicious lesions)
远处病灶的评估 (Evaluation of distant lesions)	1）体格检查 (physical examination) 2）胸部CT (Chest CT) 3）腹部 ± 盆腔影像学检查 (Abdominal ± pelvic imaging examination) 4）骨放射性核素扫描 (Bone radionuclide scanning) 5）PET-CT

For the characteristics of table data, the article designs the table extraction model TEM (Table Extraction Model), TEM contains: 1) a named entity recognition model (Bert + BiLSTM + CRF [36]) to realize named entity recognition of the table's title and content; 2) a text matching model: the article models the table title relation extraction as a text matching task, that is, the table title text semantic matching with all relation texts (Bert + ESIM [37, 38]), and the one with the highest matching degree is used as the relationship between entities. Specifically, the article implements a named entity recognition model based on Bert + BiLSTM + CRF [36], and the main framework of the algorithm is shown in Figure 5. Among them, Bert is a large-scale pre-trained language model [39, 40], which is based on a multilayer bidirectional Transformer model with sequence Mask Language Model (MLM) and Next Sentence Prediction (NSP) as the main tasks for training based on large-scale unsupervised data, which enables the Bert model to learn the semantic representation of text through a large amount of unsupervised data. In case of insufficient training tasks, the robustness of the model can be improved in downstream tasks. The self-attention mechanism is the core improvement of the Transformer model, which can give different weights to different words to get better character and sentence representations when learning sentence representations. The calculation method of self-attention is shown in Eq (3.1).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \tag{3.1}$$

In this paper, we use the Bert + ESIM based on semantic matching for relation extraction, where ESIM is an interactive text-matching method, which considers the content of the text to be matched to improve the effect of semantic similarity computation through the attention mechanism [41] in the learning text representation.

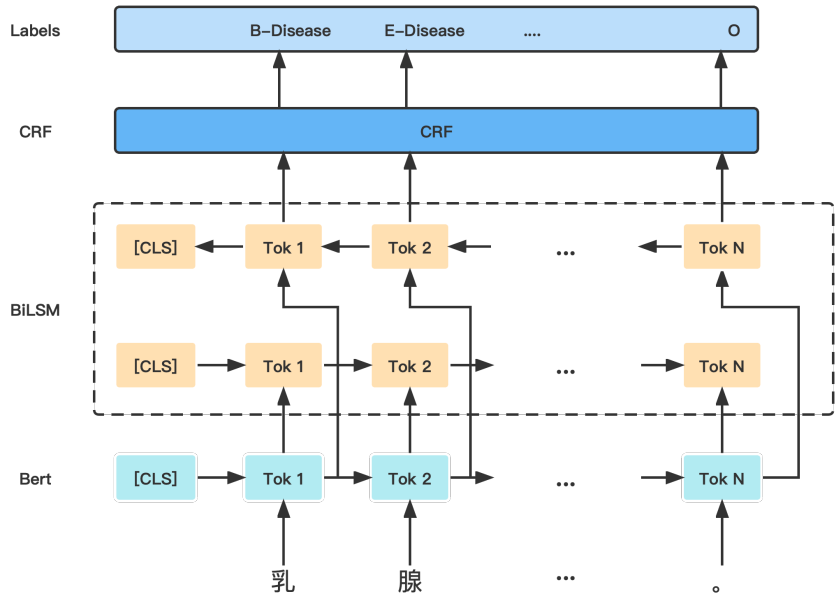

Figure 5. The illustration of Bert + BiLSTM + CRF.

3.2.4. Medical encyclopedic knowledge extraction

Medical encyclopedias are open knowledge created and maintained by medical experts, with high data quality, and are important data sources for constructing specialized disease knowledge graphs. This paper employs YouLai Doctor [I] and Medical Encyclopedia [**] as two typical encyclopedic data for extraction. Knowledge from different encyclopedia data sources can corroborate each other, and the knowledge appearing in multiple data sources is highly credible, which reduces the workload of manual review. Encyclopedias usually contain two types of knowledge: structured Infobox, and semi-structured textual knowledge. The structured Infobox, such as the Infobox corresponding to the breast cancer page in Figure 6, can extract the knowledge triple directly from Html by parsing technology, such as the Infobox in Figure 6 can directly extract (breast cancer, consultation department, breast surgery).

就诊科室：乳腺外科
Department: breast surgery
是否常见：是
Common: Yes
疾病别称：乳腺恶性肿瘤
Disease alias: breast cancer
常用检查：乳腺 X 射线、体格检查
Common examination: breast X-ray, physical examination
临床症状：乳房肿块、乳房疼痛
Clinical symptoms: breast lump, breast pain
常用药物：多比柔星、紫杉醇
Common drugs: doxorubicin, taxol

Figure 6. The infobox of medical encyclopedias.

[I]https://www.youlai.cn/
[**]https://www.yixue.com/

3.2.5. Semi-structured text knowledge extraction

This kind of knowledge is contained in the encyclopedia text. Unlike the text information in the clinical guideline, the encyclopedia text is semi-structured, i.e., the text contains many paragraphs, and different paragraphs represent different type of information, such as the disease page contains paragraphs of 'cause', 'prevalent population', 'symptoms', 'surgical treatment', 'consultation department', etc. The titles of these paragraphs are the corresponding relationship descriptions of the disease, and the contents cover the corresponding relationship entities.

In response to the above analysis, this paper pre-processes each paragraph in the encyclopedia, and concatenates the page title, paragraph title and paragraph content to form a long text. For example, the paragraph of 'Early symptoms' in the page of 'Breast cancer' is spliced as 'Early symptoms of breast cancer are not obvious, such as breast lumps and abnormal skin of the breast', etc. In the late stage, distant metastasis of cancer cells may occur, which may be manifested as enlarged lymph nodes in the ipsilateral armpit and multi-organ lesions in the whole body, which may directly threaten the life of patients. In addition, other symptoms such as bleeding and fluid accumulation may also occur. The article employs the joint extraction model TPC to extract knowledge from the stitched text and construct a triple of knowledge, such as (breast cancer, symptoms, distant metastasis of cancer cells), etc. The above method can extract knowledge triple from medical encyclopedia text.

3.2.6. Medical record event extraction

Patient medical records are clinical practices for diseases and are valuable for clinical research, drug development, and other applications. Unlike clinical guidelines and medical encyclopedias, medical record data usually contain multiple types of events, such as diagnosis, surgery, and drug treatment of diseases. They include event elements such as hospital, time, and symptoms. The traditional triple knowledge is not suitable for describing complex event knowledge composed of multiple elements, so we model the knowledge in the medical record as event. There are two main tasks in event extraction: trigger word recognition and event element extraction. For example, the text in Figure 7 is a consultation event, where 'complaint' is the trigger word, 'left breast pain, breast lump' is the 'symptom' element, '3 months' is the 'duration' element.

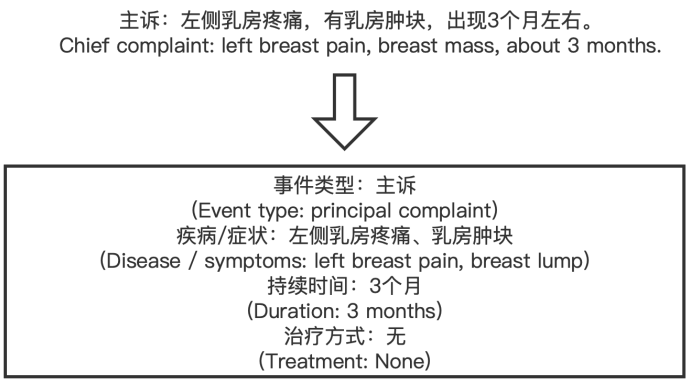

Figure 7. An example of complaint event.

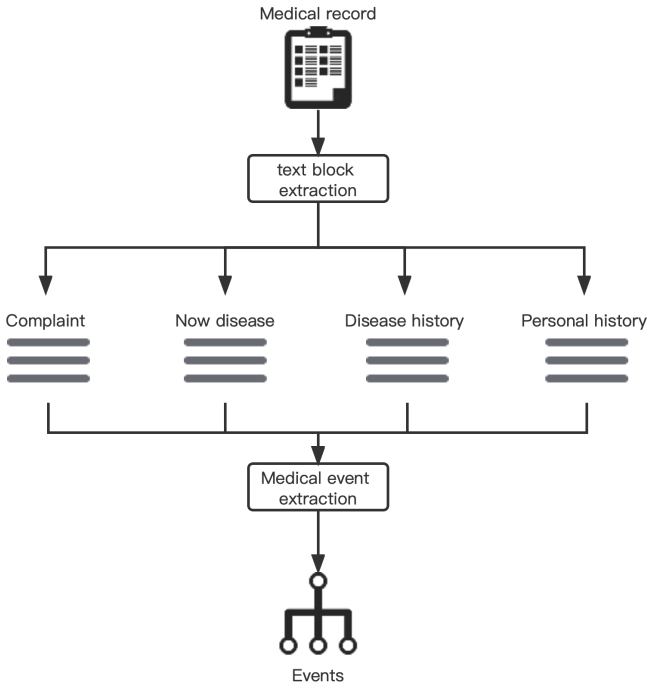

Figure 8. The framework of event extraction.

Medical record writing usually has fixed norms [42]. For example, medical records usually include 'information about the visit', 'chief complaint', 'current disease history', 'past disease history', 'allergy history', 'personal history', 'birth history', 'reproductive history', 'physical examination', etc. The different segments are usually described in different paragraphs or divided by special symbols, e.g., 'Complaint: left knee pain, lasting 3 days.'. For this characteristic of medical records, this paper model the event extraction of medical records as two steps: 1) text block extraction; 2) event extraction. The overall medical record event extraction process is shown in Figure 8. Given the medical record text, text block extraction is performed first to extract the text of fields such as 'chief complaint'. Then, the specific event knowledge is extracted by the event extraction module.

The main goal of text block extraction is to extract text blocks describing different segments of medical records from the full text of medical records, such as the text block of 'chief complaint', and the model gives the start position, end position and field type of the text block. Therefore, text block extraction is a typical sequence annotation task, and Bert + BiLSTM + CRF is employed in this paper to achieve text block extraction. The event extraction is to extract the event trigger word and event element from the result of text block extraction, such as the event of 'consultation' shown in Figure 8, which needs to extract the trigger word 'chief complaint', event element 'disease/symptom', 'ongoing event', etc. Medical record event extraction includes various types of events. In this paper, we model the medical record event extraction as a generative model (EventGen), which directly generates the corresponding structured event information by inputting textual information, and contains two main parts: classification model and event generation model. The classification model is implemented based on Bert + MLP, which identifies the event type of the input text and serves as the input to the generation

model. The event generation model is based on MedBart, a medical domain generation model trained based on Bart [43] model, and an example of 'medical visit event' extraction is given in Figure 9. Through the above two steps, the model can automatically identify the event type corresponding to the text and constrain the event generation model with the event Schema to improve the accuracy of event generation.

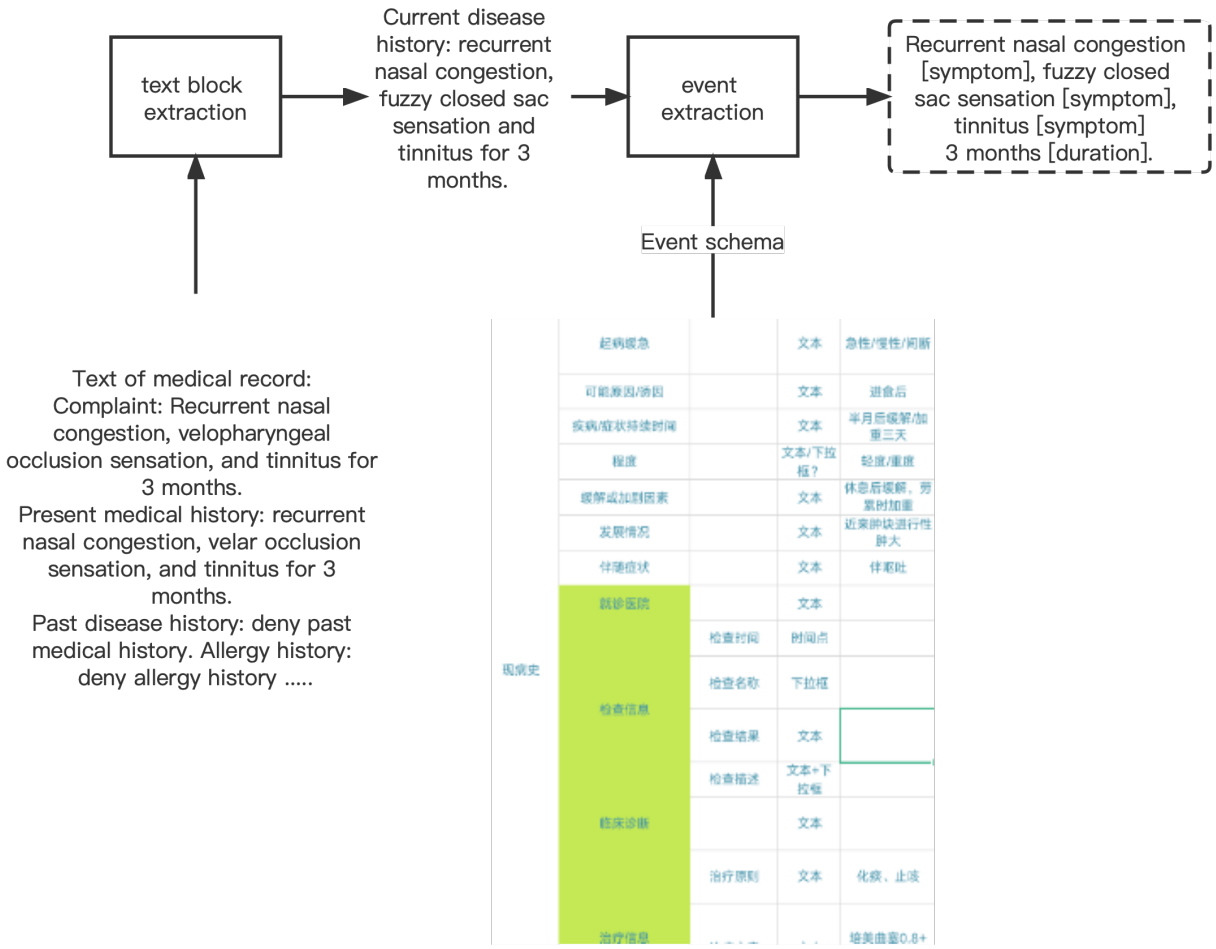

Figure 9. An example of event extraction.

3.3. Knowledge fusion

Through knowledge extraction, the article extracts triple knowledge and event knowledge from clinical guidelines, a medical encyclopedia and medical records, respectively. The knowledge extracted from different data sources has different structures (ontology level) and entity names (entity level). Thus knowledge fusion is needed to realize the fusion of data from different sources to construct a large-scale knowledge graph. Knowledge fusion contains two levels of fusion: ontology fusion and entity fusion. Ontology fusion refers to the fusion at the level of entity type, relation type and attribute type, e.g., 'symptom' in encyclopedia and 'clinical observation' in guidelines can be mapped to the same entity type, and 'causative factor' in guidelines can be mapped to 'disease' can be mapped to

'etiology'. Entity fusion is fusion at the entity level, e.g., 'ductal carcinoma in situ' and 'non-invasive breast cancer' refer to the same disease, but have different names and need to be linked, also known as medical terminology normalization [44]. Since ontology fusion involves fewer data and affects many entities and triple knowledge, the paper employs a manual collation and mapping approach developed with the participation of medical experts and knowledge graph experts.

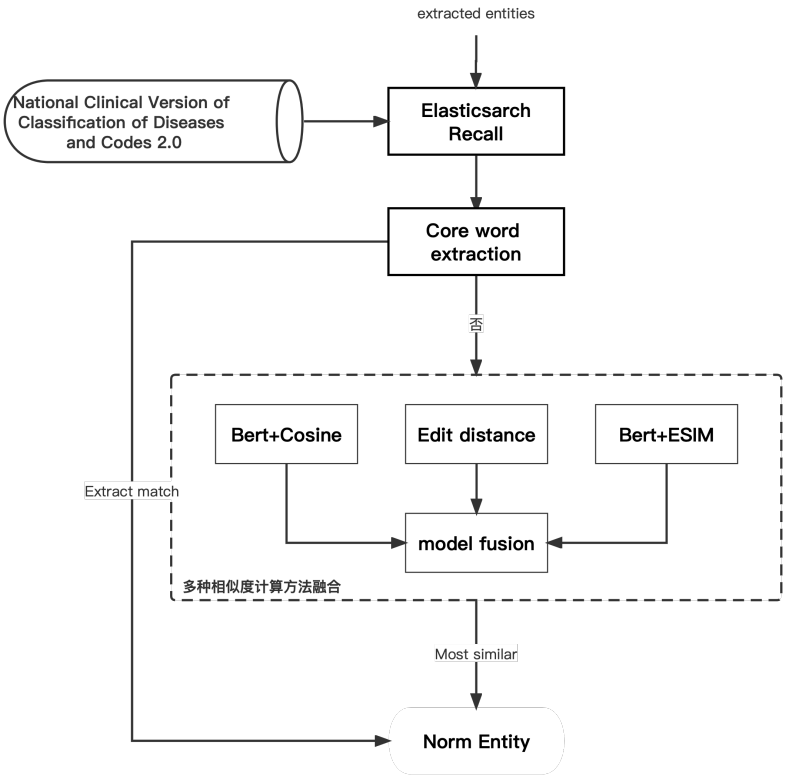

Figure 10. The framework of medical term normalization.

Entity fusion is the different entity names expressing the same concept at the entity level, which has led to multiple names for many medical concepts because healthcare systems in different regions are not interconnected. Medical terminology normalization is also key to constructing a large-scale medical and health intelligence knowledge graph. In this paper, we employ the Chinese National Clinical Version of Disease Classification and Coding 2.0 as a standard terminology base, and map entities extracted from different data sources to the standard terminology base. The following problems exist in the normalization of medical terminology: 1) large size of the standard library: the National Clinical Version of Classification of Diseases and Codes 2.0 contains about 40,000 disease terms, plus more than 100,000 standard terms such as examinations and drugs; 2) different entities with similar wording: there are a large number of entities with a similar wording but not the same in medical terminologies, such as 'parametrial breast cancer' and 'left breast cancer' have a high degree of literal similarity, but 'parametrial breast cancer' does not even belong to breast cancer, so they are completely different entities; 3) the names of the same entity may have a large number of literal differences: there are also a large number of medical terms that have completely different literal but refer to the same disease.

For example, 'Alzheimer's disease' and 'A er zi hai mo zheng' are the same disease, but they are entirely different'. To address these characteristics of medical named entity names, the paper designs a three-stage entity resolution (TSER) algorithm, the overall framework of which is shown in Figure 10.

The first stage is term recall: we first recall top 100 candidate entities from the terms base by TFIDF-based text matching, this stage can significantly improve the speed of entity fusion and cope with problem 1).

The second stage is core word extraction: medical terms usually consist of multiple parts, such as the disease term 'left breast carcinoma in situ', which contains the orientation word 'left', the body word 'breast, the nature word 'in situ'. The term 'in situ' and the disease term 'cancer' are usually composed of several parts. The importance of different types of words to the terminology varies, among which disease words, body words and nature words are more important to determine the disease. Therefore, in this paper, the terms are split based on component dictionaries, and then different components are given different weights. In this stage, different disease terms can be distinguished by core words, such as 'invasive breast cancer' and 'invasive pancreatic cancer', and by body terms 'breast' and 'pancrea' to address the challenge 2).

Phase 3 is semantic matching: This phase obtains the core word split candidates based on the above recall and core word extraction modules and then employs various types of semantic similarity calculation methods, including edit distance-based algorithms (literal similarity), similarity calculation methods based on independent representation learning (Bert + Cosine) and similarity calculation methods based on interactive representation learning (Bert + ESIM), where Bert+Cosine and Bert + ESIM mainly perform similarity computation from the semantic level to address the challenge 3). The three similarities are finally fused by Eq (2) to obtain the final similarity and then ranked from highest to lowest similarity. Where w1, w2 and w3 are learnable parameters, and the final results are obtained by model training.

The main processes of breast cancer knowledge graph construction include 1) extracting triples from guideline text and encyclopedia text using the joint extraction model (TPC); 2) extracting triples from guideline tables using the table extraction method (TEM); 3) extracting event knowledge from medical records using the event generation model (EventGen); 4) the data from different sources are aligned and correlated based on the entity fusion model (TSER) to merge the triples and align the entities in the events to construct the breast cancer knowledge graph.

4. Breast cancer knowledge graph construction and evaluation

4.1. Data pre-processing

The data sources for constructing the breast cancer knowledge graph mainly include: clinical guidelines, medical encyclopedias and medical records. In this paper, we first employ LabelStudio to build a data labelling system for data labelling, and a labelling example is shown in Figure 11. In order to improve the accuracy of the labeled data, this paper adopts an interactive manual and model labeling method, 1) that is, firstly using the model for pre-labeling; 2) then using manual for proofreading, then using a new batch of data for pre-labeling and then manual proofreading, by this way labeling joint extraction, table extraction, and event extraction models. The specific size of the labelled data is shown in Table 2.

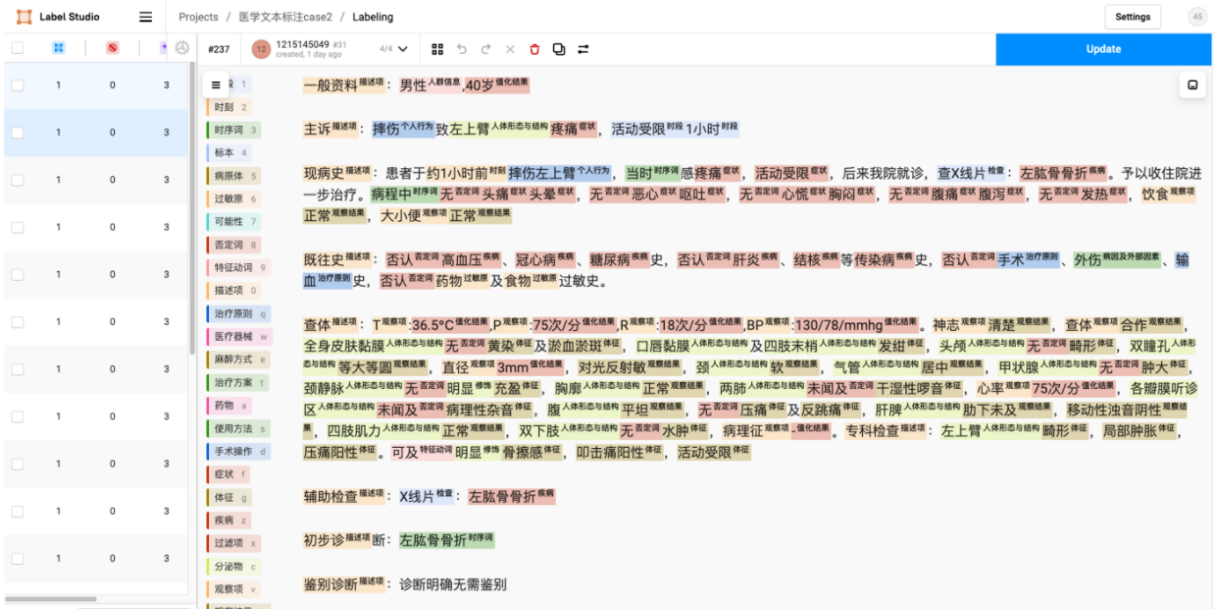

Figure 11. An example of interface of label system.

Table 2. The annotated dataset.

Type	Number of Annotated data	Source
Joint learning	5000 sentences	Guidelines
Tabulur exaction	100	Guidelines
Named entity recognition	20,000 sentences	Guidelines

4.2. Experiment settings

In this paper, 90% of the labelled data of each data is randomly selected as the training set, and the remaining 10% of the data is used as the test data. In this paper, we implement the relevant models based on Pytorch where the Bert model is MCBert [††] and fine-tuning based on our own medical literature and forum text corpus, and implement TPC, Bert + BiLSTM + CRF, Bert + ESIM, EventGen and other models based on Transformers framework. The hidden size of BiLSTM is set 200, the dimension of MLP is set at 200, batch_size is set to 128, train_step is set to 10,000, learning_rate is set to 0.001, dropout is set to 0.2 and the number of layers of BiLSTM is set to 4.

4.3. Knowledge extraction

Based on the annotated data, this paper trains the joint extraction model, guideline table extraction model, encyclopedia entity extraction model, medical record text block extraction, and medical record event extraction models. In this paper, we extracted entities, relations, triples, and events from all the clinical guidelines, medical encyclopedia texts, and medical records, and realized the fusion of triples from different data sources through knowledge fusion, and finally constructed a breast cancer knowledge graph. The scale of the knowledge graph extracted by TPC, TEM and the fused KG is

[††]https://github.com/alibaba-research/ChineseBLUE

shown in Table 3, and Neo4j [‡‡] is employed for knowledge storage in this paper.

Table 3. The statistic of breast cancer knowledge graph.

Data Type	TPC	TEM	Fusion
Entity	8734	6542	13,213
Relation	64	64	64
Triple	118,765	67,523	132,454

4.4. Knowledge graph evaluation

4.4.1. Triple knowledge

The triple and event knowledge in the breast cancer knowledge graph can directly provide intelligent services, such as knowledge answering and clinical decision support system. The quality of the knowledge graph has an essential impact on the above applications. In this paper, we employ manual to evaluate the extracted knowledge, and finally evaluate the extracted triple and events using the accuracy rate, which is calculated as shown in Eq (4.1), where N_{right} is the number of correct triples/events and N is the number of all triples/events. Specifically, in this paper, 2000 triple and 1000 events are randomly selected, and two medical experts can label whether the corresponding knowledge is correct. A third doctor is used to label when there is inconsistent labeling, and the principle of the minority following the majority is adopted as the final labeling. In order to better verify the effectiveness of model extraction, we implement the baseline model of entity extraction and relation extraction based on dictionary matching (Dictionary) and pipeline-based entity extraction (Bert + BiLSTM + CRF) and relationship recognition (Bert + MLP classification) for training in triple knowledge extraction. The correct rates are shown in Table 4. The comparison shows that the method TPC proposed in the paper can significantly improve the accuracy rate of triple extraction, avoid the problem of error propagation through the joint learning method, and better utilize the relationship information to optimize the result of entity recognition to achieve better triple extraction performance.

$$Accuracy = N_{right}/N \tag{4.1}$$

Table 4. The accuracy of triple extraction.

Model	Accuracy
Dictionary	67.8
Pipeline	82.3
Our	93.2

4.4.2. Event knowledge

The comparison models for event extraction are TANL and Multi-task TANL using accuracy (Precision), recall (Recall) and F1-value (F1-value) as evaluation metrics, and the results of event extraction are shown in Figure 12. The results of knowledge extraction show that the method proposed in this

[‡‡]https://neo4j.com/

paper can construct high-quality breast cancer-specific knowledge graphs with better knots in accuracy, recall and F1-value, and can extract more high-quality triples from the same data source.

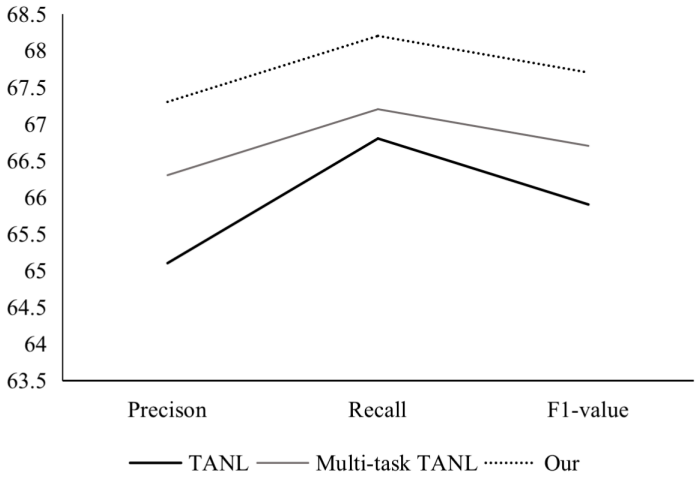

Figure 12. An example of interface of label system.

4.4.3. Knowledge fusion

This paper employs a lexicon-based approach (Dictionary) and a Bert-based approach for representing cosine similarity as baseline models. This paper uses Top@N as an evaluation metric, i.e., the proportion of the top N results of matching that contain correct results. All three models first recall the top 100 candidate terms using Elasticsearch, and then perform refined similarity calculations using the corresponding models. The results of term normalization are shown in Table 5. The results show that TSER can achieve better results on Top@1, Top@5, and Top@10 results and better knowledge fusion.

Table 5. The result of term normalization.

Model	Top@1	Top@5	Top@10
Dictionary	62.1	75.1	84.1
Bert + Cosine	81.1	86.3	89.2
Bert + ESIM	84.3	88.6	92.1

5. Applications

The breast cancer knowledge graph has direct application value for breast cancer knowledge question answering, clinical diagnosis and treatment, drug development, follow-up and health management. In this paper, we introduce the application of the breast cancer knowledge graph constructed in this paper with two applications: medical question answering and medical record search.

5.1. Medical knowledge-based question answering

Knowledge graph-based question answering (KBQA) is a typical application of knowledge graph, and this paper implements question answering requirements related to breast cancer treatment and

health management. In this paper, question answering is modelled as two steps: question understanding and knowledge query. Question understanding mainly consists of entity recognition and intention understanding, and knowledge query is transcribed into Sparql query based on question understanding, and is available on Neo4j to get the knowledge of the user's query. Question understanding refers to understanding the user's query, which mainly includes entity recognition and intent understanding, where entity recognition is reused for entity recognition based on Bert + BiLSTM + CRF trained in Section 3.1. Intent recognition determines the relationship, attribute or path information [45] corresponding to the user's question in the knowledge graph. Due to the lack of training data, this paper models intent recognition as a text-matching task to achieve a cold start. Because there are many text matching datasets are released. Specifically, in this paper, all relations, attributes and common paths are used as candidate intents, and for each input question, the entity is rewritten to the entity corresponding type after entity recognition, and then the intent with the closest semantic similarity to the candidate intent is obtained as the user intent by semantic computation, and the Bert + ESIM algorithm is used as the semantic matching computation model. In this way, identifying entities and intents in the interrogative sentence is realized, and the corresponding entities and intents are transformed into Sparql query statements utilizing rules to realize knowledge-based question answering. In order to verify the effectiveness of the method, this paper uses the matching method based on edit distance (Edit), the method based on Bert representation of similarity as the baseline for intent understanding. We employ top@N as the calculation method for model accuracy, i.e., the proportion of the top N answers containing the target answer. In this paper, 500 questions were written manually to cover all relationship/attribute types. For example, the question is 'What medicine can treat breast cancer' and the answer is 'Drugs for breast cancer include paclitaxel, doxorubicin, ...'. The final results are shown in Figure 13. The results show that, in the absence of training data, the method proposed in this paper can better achieve the cold start of knowledge question answering and make the breast cancer knowledge graph work better.

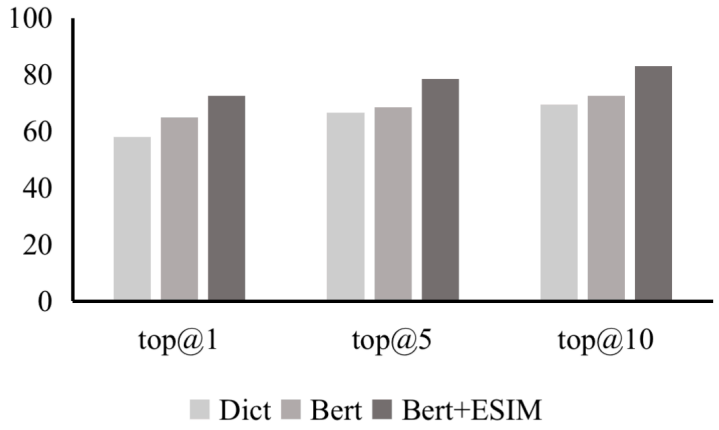

Figure 13. The results of KBQA.

5.2. Medical record retrieval

When doctors analyze and make diagnostic decisions about a patient's condition, they can often refer to historical medical records in medical record databases to see how patients with similar conditions have been treated and their treatment outcomes. Therefore, precisely finding similar medical records

from large-scale medical record databases can improve treatment and alleviate problems such as the uneven distribution of medical resources in China. Traditional medical record matching is mainly based on string matching through text information in medical records, however, different regions, hospitals, and doctors have different habits in writing medical records, and the description of the same condition varies greatly. Some physicians describe the symptoms in more detail, such as 'swelling and pain in the right knee for about 3 days', while others describe the condition more concisely, such as 'right knee pain 3 days'. These two descriptions express the same symptom. However, the literal difference may be large, and the matching degree is low based on the character-matching method. This paper employs the results of medical record event extraction to perform structured matching. As both descriptions mentioned above can be extracted as diagnostic events, the events' elements are identical. Matching by events can better utilize the essential semantic information in medical records and reduce the problem of inaccurate matching due to linguistic diversity. The similarity between different diseases can be calculated based on term normalization and disease knowledge graph. For example, the similarity between 'left breast cancer' and 'breast cancer' is greater than that between 'parametrial breast cancer' and 'breast cancer'. However, calculating the literal similarity alone may give the opposite result. The disease knowledge graph contains 'IsA' relationship and a pendant class relationship, which can be used to determine the semantic similarity between entities by calculating the shortest path between two nodes, thus improving the accuracy of terminological similarity calculation in medical record retrieval. This paper constructs a medical record database containing 20,000 medical records after desensitization based on the data accumulated in the hospital. It performs event extraction for each medical record through the medical record event extraction method. Elasticsearch stores medical record text and medical record events, respectively. From them, 100 medical records are randomly selected, and their texts are manually retrieved to form 100 new medical records with different description forms. Using these 100 new medical records as test data, medical record retrieval is performed to determine whether the original medical records can be found, and Top@N is used as the evaluation criterion for accuracy. In this paper, we use a text-based retrieval approach (Text-Based) and an event-matching approach (Event-Based) for retrieval to find the most relevant medical records. Specifically, full-text retrieval was used for the text-based retrieval, and structured retrieval was used for the event-based retrieval, i.e., matching event types and event elements, respectively. The results of the medical record retrieval are shown in Table 6. The results show that event-based medical record retrieval significantly outperforms text-based retrieval, mainly because event extraction can better handle the problem of linguistic diversity by directly using critical information for structured matching.

Table 6. The result of medical record retrieve.

Model	Top@1	Top@5	Top@10
Textbased	45.1	53.2	58.1
Eventbased	71.2	78.1	80.1

6. Conclusions

In this paper, we realize knowledge extraction from multiple heterogeneous data sources based on information extraction, knowledge graph, and event extraction to construct the breast cancer knowledge graph. The experimental results show that our method can construct high-quality knowledge graphs.

The example analysis shows that the breast cancer knowledge graph can also better support the intelligence of breast cancer diagnosis and treatment scenarios such as question answering and medical record retrieval. In the future, we plan to integrate the information from academic literature and development reports into the knowledge graph to form a complete knowledge graph.

Acknowledgments

This work is supported by the Foundation of BAAI under Grant no. BAAI2021CXZX04 and the National Natural Science Foundation of China under Grants no. 62076233 and the National Social Science Foundation of China under Grant no. 22BTQ010 and Major innovation project of Chinese Academy of Social Sciences under Grants no. 21ZD304.

Conflict of interest

The authors declares there is no conflict of interest.

References

1. X. Zou, A survey on application of knowledge graph, *J. Phys. Conf. Ser.*, **1487** (2020), 12016. https://doi.org/10.1088/1742-6596/1487/1/012016
2. M. Kejriwal, Knowledge graphs and COVID-19: opportunities, challenges, and implementation, *Harv. Data Sci. Rev.*, **11** (2020), 300.
3. Q. H. Nguyen, T. T. Do, Y. Wang, S. S. Heng, K. Chen, W. H. M. Ang, et al., Breast cancer prediction using feature selection and ensemble voting, in *2019 International Conference on System Science and Engineering (ICSSE)*, IEEE, (2019), 250–254.
4. K. Zhang, X. Ren, L. Zhuang, H. Zan, W. Zhang, Z. Sui, Construction of chinese medicine knowledge base, in *Workshop on Chinese Lexical Semantics*, Springer, (2020), 665–675. https://doi.org/10.1007/978-3-030-81197-6_56
5. P. H. Martins, Z. Marinho, A. Martins, Joint learning of named entity recognition and entity linking, preprint, arXiv:1907.08243.
6. J. Noh, R. Kavuluru, Joint learning for biomedical ner and entity normalization: encoding schemes, counterfactual examples, and zero-shot evaluation, in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, (2021), 1–10.
7. L. Liu, M. Wang, M. Zhang, L. Qing, X. He, Uamner: uncertainty-aware multimodal named entity recognition in social media posts, *Appl. Intell.*, **52** (2022), 4109–4125. https://doi.org/10.1007/s10489-021-02546-5
8. S. S. Paliwal, D. Vishwanath, R. Rahul, M. Sharma, L. Vig, Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images, in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, (2019), 128–133.
9. W. Xiang, B. Wang, A survey of event extraction from text, *IEEE Access*, **7** (2019), 173111–173137. https://doi.org/10.1109/ACCESS.2019.2956831

10. Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, et al., Unified structure generation for universal information extraction, preprint, arXiv:2203.12277.

11. B. P. Nguyen, H. N. Pham, H. Tran, N. Nghiem, Q. H. Nguyen, T. T. Do, et al., Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records, *Comput. Methods Programs Biomed.*, **182** (2019), 105055. https://doi.org/10.1016/j.cmpb.2019.105055

12. X. Zhao, Y. Jia, A. Li, R. Jiang, Y. Song, Multi-source knowledge fusion: a survey, *World Wide Web*, **23** (2020), 2567–2592. https://doi.org/10.1007/s11280-020-00811-0

13. A. Hogan, E. Blomqvist, M. Cochez, C. D'Amato, G. D. Melo, C. Gutierrez, et al., Knowledge graphs, *ACM Comput. Surv.*, **54** (2021), 1–37. https://doi.org/10.1145/3466817

14. M. Wang, X. He, L. Liu, L. Qing, H. Chen, Y. Liu, et al., Medical visual question answering based on question-type reasoning and semantic space constraint, *Artif. Intell. Med.*, **131** (2022), 102346. https://doi.org/10.1016/j.artmed.2022.102346

15. X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, et al., Multi-modal knowledge graph construction and application: A survey, preprint, arXiv:2202.05786.

16. L. Liu, M. Wang, X. He, L. Qing, H. Chen, Fact-based visual question answering via dual-process system, *Knowl. Based Syst.*, **237** (2022), 107650. https://doi.org/10.1016/j.knosys.2021.107650

17. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, T. M. Mitchell, Toward an architecture for never-ending language learning, in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, **24** (2010), 1306–1313.

18. D. Vrandečić, Wikidata: A new platform for collaborative data collection, in *Proceedings of the 21st International Conference on World Wide Web*, (2012), 1063–1064.

19. L. Liu, M. Wang, X. He, L. Qing, J. Zhang, Extracting relational facts based on hybrid syntax-guided transformer and pointer network, *J. Intell. Fuzzy Syst.*, **40** (2021), 12167–12183. https://doi.org/10.3233/JIFS-210281

20. H. Lv, H. Liang, F. Ma, Constructing knowledge graph for financial equities, *Data Anal. Knowl. Discovery*, **4** (2020), 27–37.

21. F. Sovrano, M. Palmirani, F. Vitali, Legal knowledge extraction for knowledge graph based question-answering, in *Legal Knowledge and Information Systems*, IOS Press, (2020), 143–153.

22. Y. Wei, H. Wang, J. Zhao, Y. Liu, Y. Zhang, B. Wu, Gelaigelai: a visual platform for analysis of classical chinese poetry based on knowledge graph, in *2020 IEEE International Conference on Knowledge Graph (ICKG)*, IEEE, (2020), 513–520.

23. F. Gong, M. Wang, H. Wang, S. Wang, M. Liu, Smr: medical knowledge graph embedding for safe medicine recommendation, *Big Data Res.*, **23** (2021), 100174. https://doi.org/10.1016/j.bdr.2020.100174

24. H. Chen, N. Hu, G. Qi, H. Wang, Z. Bi, J. Li, et al., Openkg chain: A blockchain infrastructure for open knowledge graphs, *Data Intell.*, **3** (2021), 205–227.

25. A. Chatterjee, C. Nardi, C. Oberije, P. Lambin, Knowledge graphs for COVID-19: An exploratory review of the current landscape, *J. Pers. Med.*, **11** (2021), 300. https://doi.org/10.3390/jpm11040300

26. S. Ji, S. Pan, E. Cambria, P. Marttinen, S. Y. Philip, A survey on knowledge graphs: Representation, acquisition, and applications, *IEEE Trans. Neural Networks Learn. Syst.*, **33** (2021), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843
27. B. Xie, S. Li, F. Lv, C. H. Liu, G. Wang, D. Wu, A collaborative alignment framework of transferable knowledge extraction for unsupervised domain adaptation, *IEEE Trans. Knowl. Data Eng.*, **2022** (2022). https://doi.org/10.1109/TKDE.2022.3185233
28. J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, *IEEE Trans. Knowl. Data Eng.*, **34** (2020), 50–70. https://doi.org/10.1007/s10618-019-00656-w
29. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, (2017), 30.
30. S. Edunov, A. Baevski, M. Auli, Pre-trained language model representations for language generation, preprint, arXiv:1903.09722.
31. L. X. Liang, L. Lin, E. Lin, W. S. Wen, G. Y. Huang, A joint learning model to extract entities and relations for chinese literature based on self-attention, *Mathematics*, **10** (2022), 2216. https://doi.org/10.3390/math10132216
32. M. Zhang, Y. Chen, J. Lin, A privacy-preserving optimization of neighborhood-based recommendation for medical-aided diagnosis and treatment, *IEEE Internet Things J.*, **8** (2021), 10830–10842. https://doi.org/10.1109/JIOT.2021.3051060
33. B. An, X. Han, C. Fu, L. Sun, Retrofitting soft rules for knowledge representation learning, *Big Data Res.*, **24** (2021), 100156. https://doi.org/10.1016/j.bdr.2020.100156
34. J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, et al., The evolution of protégé: an environment for knowledge-based systems development, *Int. J. Human Comput. Stud.*, **58** (2003), 89–123. https://doi.org/10.1016/S0031-9406(05)60588-3
35. M. Peleg, Computer-interpretable clinical guidelines: a methodological review, *J. Biomed. Inf.*, **46** (2013), 744–763. https://doi.org/10.1016/j.jbi.2013.06.009
36. Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, X. Bai, Named entity recognition using bert bilstm crf for Chinese electronic health records, in *2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (cisp-bmei)*, IEEE, (2019), 1–5.
37. Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, K. K. Ma, Esim: Edge similarity for screen content image quality assessment, *IEEE Trans. Image Process.*, **26** (2017), 4818–4831. https://doi.org/10.1109/TIP.2017.2718185
38. L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, Bert-attack: Adversarial attack against bert using bert, preprint, arXiv:2004.09984.
39. E. K. W. Leow, B. P. Nguyen, M. C. H. Chua, Robo-advisor using genetic algorithm and bert sentiments from tweets for hybrid portfolio optimisation, *Expert Syst. Appl.*, **179** (2021), 115060. https://doi.org/10.1016/j.eswa.2021.115060
40. T. Nguyen-Vo, Q. H. Trinh, L. Nguyen, T. T. Do, M. C. H. Chua, B. P. Nguyen, Predicting antimalarial activity in natural products using pretrained bidirectional encoder representations from transformers, *J. Chem. Inf. Model.*, **62** (2021), 5050–5058. https://doi.org/10.1021/acs.jcim.1c00584

41. Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, *Neurocomputing*, **452** (2021), 48–62. https://doi.org/10.1007/s43830-021-0173-9

42. A. E. Patanwala, A practical guide to conducting and writing medical record review studies, *Am. J. Health Syst. Pharm.*, **74** (2017), 1853–1864. https://doi.org/10.2146/ajhp170183

43. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, et al., Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, preprint, arXiv:1910.13461.

44. Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, S. Yu, Coder: Knowledge-infused cross-lingual medical term embedding for term normalization, *J. Biomed. Inf.*, **126** (2022), 103983. https://doi.org/10.1016/j.jbi.2021.103983

45. Y. Shen, N. Ding, H. T. Zheng, Y. Li, M. Yang, Modeling relation paths for knowledge graph completion, *IEEE Trans. Knowl. Data Eng.*, **33** (2020), 3607–3617. https://doi.org/10.1109/TKDE.2020.2970044

AIMS Press