Int. j. inf. tecnol. (January 2023) 15(1):53–65
https://doi.org/10.1007/s41870-022-01145-y

ORIGINAL RESEARCH

Knowledge graph enrichment from clinical narratives using NLP, NER, and biomedical ontologies for healthcare applications

Anjali Thukral[1] · Shivani Dhiman[2] · Ravi Meher[3] · Punam Bedi[2]

Received: 20 June 2022 / Accepted: 13 December 2022 / Published online: 3 January 2023

Abstract Electronic health records (EHR) contain patients' health information in varied formats such as clinical reports written in natural language, X-rays, MRI, case/discharge-summary, etc. One of its essential constituents is clinical narratives which contain significant clinical findings of a patient. Since the clinical narratives are stored in natural language, clinical evidence, significant findings, and other observations written in it by doctors remain locked. This free text information is incomprehensible by machines while processing EHRs in healthcare applications such as Clinical Decision Support System. The proposed work, Clinical Narratives to Knowledge Graph (N2K) Mapper algorithm, is an effort to map clinical narratives to a Knowledge Graph (KG) so that important clinical details as recommended by doctors can be used effectively by healthcare applications. The KG is defined by an ontological semantic meta-data called Healthcare Ontology. The N2K Mapper algorithm uses natural language processing to parse the clinical narratives and recognizes medicinal vocabulary using named entity recognition and various existing biomedical ontologies. This semantically retrieved information is then mapped onto a KG. Besides enriching the KG from clinical narratives, the algorithm also augments the biomedical ontologies with new medicinal vocabulary leading to sustainable semantic structure for future references. An experimental study was performed on Chest X-ray radiology reports taken from the Medical Information Mart for Intensive Care (MIMIC)-III dataset. It showed that the N2K Mapper was able to find and recognize approximately 80% of the total identified entities in the existing biomedical ontologies. The remaining 20% were augmented in the biomedical ontologies with an expert's assistance. The N2K Mapper showed an accuracy of 81.5%, precision of 78.5%, recall of 85.07%, and F1-score of 78.97%.

Keywords Electronic health record · Clinical narratives · Natural language processing · Ontology · Named entity recognition · Knowledge graph

✉ Shivani Dhiman
shivani@cs.du.ac.in

Anjali Thukral
athukral@keshav.du.ac.in

Ravi Meher
ravimeher@gmail.com

Punam Bedi
pbedi@cs.du.ac.in

[1] Keshav Mahavidyalaya, University of Delhi, Delhi 110034, India

[2] Department of Computer Science, University of Delhi, Delhi 110007, India

[3] Maulana Azad Medical College, University of Delhi, Delhi 110002, India

1 Introduction

The researchers at the University of Groningen and the University of Twente were the first to coin the term 'KG' in the 1980s [1]. However, it essentially got a wider acceptance in 2012 after a mention of the 'launch of KG' in one of Google's blogs [2]. Since then, a hike has been seen in the literature [3, 4]. The application of KG does not limit to any one particular domain, and it can be seen in the education sector, agriculture, social media, networking, and many more [3, 4].

The reach of KG has been extended to the medicinal domain as well [5, 6]. KGs can semantically represent patients' EHR which are voluminous and available in various formats. Valid and time-bound inferences can also be

inferred from the medicinal knowledge stored in KG. Many intelligent medical applications software for Disease Diagnosis and CDSS can be built based on the semantic knowledge extracted from these KGs, which otherwise is difficult due to the unstructured data in EHRs.

One of the crucial documents in EHR is Clinical notes or Clinical narratives, which contain critical radiological or/and pathological findings related to a patient. These pathological findings play an important role in clinical decision-making and disease diagnosis. Clinical narratives are usually written in natural language and accompanied by other statistical data. The 'narratives' are generally kept as it is in EHRs. As a result, the artificial intelligence (AI) based clinical application software cannot capture important findings from these narratives. However, if clinical narratives are represented in a format understandable by AI applications, then a better diagnosis or decision can be obtained. Therefore, it becomes more important to translate clinical narratives effectively while retaining the medicinal vocabulary and semantics.

KGs provide a solution for such semantic representation of information [3]. The semantically mapped KGs can be used to produce valuable insights [7]. This paper proposes a N2K Mapper algorithm that makes the clinical narratives machine-understandable. It converts the clinical narratives to KG using NLP, NER, and biomedical ontologies. N2K Mapper extracts clinical findings from the clinical narratives to map those to a KG. It can be used in several AI-based applications in healthcare services, including recommendations on preventive healthcare measures and differential disease diagnosis.

The contributions of this paper are detailed as follows:

- The paper proposes a framework to map patients' clinical narratives (such as radiology reports) into semantically structured and machine-understandable KGs.
- HO, a conceptual schema designed to incorporate a generic EHR, forms the definition for the KG.
- The framework uses the N2K Mapper algorithm which
    - maps the semantically retrieved entities from clinical narrative and form triples by relating extracted entities using HO and subsequently enriching the KG.
    - It also augments the biomedical ontologies with new medicinal vocabulary leading to sustainable semantic structure for future references.

The paper is organized into four sections: Sect. 2 discusses the related work. Section 3 describes the framework to enrich the KG from clinical narratives using the proposed N2K Mapper. An experimental study of N2K Mapper using the MIMIC-III dataset is demonstrated in Sect. 4. The results obtained in the experimental study are presented in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related work

The following steps are required to make clinical records, particularly radiology narratives machine-understandable: Retrieval of important information from unstructured data, mapping, and augmentation of the extracted information to the knowledge base. Therefore, this section explores two significant areas related to the proposed work: (1) techniques used to extract patient-related information from clinical reports, and (2) KG representation.

2.1 Techniques used to extract patient-related information from clinical reports

Clinical narratives contain text in natural language. Various methods are used to analyze this text and make it machine readable. The authors in [8] have used dictionary-based term recognition (MetaMap) and automatic term recognition (FlexiTerm) strategies to recognize medical terminologies from MRI reports of knee injuries. They have used manual annotations afterwards to include the residual terms left by the techniques above-mentioned and organize these terms in an ontology. In [9], biomedical NER (Bio-NER), a deep neural network-based method, was proposed to classify the biomedical entities in their respective labels. In Bio-NER, convolutional neural network (CNN) was used to extract the word and sentence-level features from clinical narratives. The authors in [10] extracted names of disease and symptom concepts from unstructured notes of EHR. Two resources were used for concept information: the Unified Medical Language System (UMLS) dictionary and Google's Health KG (GHKG). The concepts in EHR were string matched with the alias and synonyms of concepts present in UMLS and GHKG. These extracted concepts are mapped to International Classification of Diseases (ICD)-9 codes. In [11], NER was used to extract entities from clinical notes using Bidirectional Encoder Representations from Transformers (BERT) variants, including BioBERT, BioClinicalBERT, Bio-DischargeSummury, and BioRoberta. Also, the relation between extracted entities was established using BERT variants only. In [12], the n-grams strategy was used to extract information from narratives. UMLS was used to recognize medical terms from the extracted n-grams from narratives.

2.2 KG representation

The extracted information must be stored and organized semantically so that AI techniques can be applied to it and new knowledge may be inferred. Various data representation techniques have been discussed in literature where clinical narratives are used as input data. In [8], the information extracted with a dictionary-based method and FlexiTerm was organized into a template with a predefined frame slot

for a particular finding. The template provides the structure to the extracted information and can ease the retrieval of it. In [13], the author organizes the extracted medical entities into a KG, Elsevier Healthcare KG (HG). It was developed to identify text snippets from medical literature for trusted medical decision support systems. HG was used as a source of triples that were utilized to identify matching sentences from the text snippets. HG can be applied to extract medical relations from medical text using a given query. In [14], the authors represented electronic medical record (EMR) consisting of various medical entities into a KG. They have proposed a methodology to develop a medical KG from EMR, and graph embedding was learned from the generated KG. The paper also proposed a ranking function to consider probability, specificity, and reliability in KG. In [15], the authors proposed a system to recommend a suitable doctor for a patient in a healthcare consultation program. The system extracted interactive features between doctors and patients from the semantically represented KG. These features were fed to a deep neural network with layer-wise relevance propagation to produce explainable recommendations.

In the research [10], titled 'Learning a Health KG from EMR', the authors used KG as its knowledge representation technique. They have proposed a methodology for automatic construction of KG that finds a relationship between disease and symptoms using data available in EMR. It aims to benefit self-diagnostic symptom checkers and CDSS. In [11], the authors built a KG using BERT from clinical narratives. BERT was used for feature extraction from clinical notes followed by a conditional random field (CRF) layer to provide a link between recognized medical entities. Table 1 summarizes a comparative study of mapping clinical narratives to knowledge structure.

Ontologies are used to organize information collected from various sources, including biomedical literature, health web portals, EMR, etc., into a semantic structure. An ontological schema can depict medicinal concepts and their inter-relations. It forms the basis for the curated KG, which can be utilized in various medical applications such as disease diagnosis. The work with biomedical ontologies was first started with genetics, pathways, and proteins [22] which is now extended to medical diagnosis [18, 19]. In [18], Traditional Chinese Medicine (TCM) healthcare ontology was developed from reference books, papers, textbooks, and dictionaries. It defines the domain concepts of TCM and their relations extracted from medicinal resources. In [5], 'Conceptual KG (CKG)' was developed in which the relationships between disease concepts were depicted. The conceptual schema obtained was termed HuadingCKG and defined the events related to the concept nodes (disease). CKG was constructed using the other existing ontologies UMLS, NCBI, etc. It was developed for the cardiovascular domain. In [17], EHR was mapped to Health Data Model using the Metathesaurus of HDM providing a terminology glossary. The authors in [8] have developed the 'TRAK' ontology. It models the knowledge and conditions of knee recovery. From the literature, the most nearby ontology schema to ours is given in [17] whereas other schemas were limited to specialized body areas or diseases. The schema in [17] is centralized around patient information and mapped to ontology, similar to our work. In our paper, HO incorporates significant concepts such as demographics and radiology reports in addition to the concepts in [17] (patient entity, Diagnosis, Medical Examination (Laboratory tests), and Chief Complaint). The radiology report's findings and patient demographics significantly contribute to identifying the possible diagnosis.

In a primary investigation of disease diagnosis, symptoms are the first source of information. The literature have studies where the relations were found between symptoms and diseases from de-identified patient notes [10], from titles and abstracts of biomedical publications [19], from Chinese EMR [21]. However, relationships based on symptoms and diseases only, may lead to inaccurate diagnosis or no diagnosis at all. This may happen due to unavailability of other influencing factors or parameters such as test/investigation results and radiology reports. For example, in cardiac and pulmonary diseases, a chest X-ray is the first test recommended by doctors. Here, clinical findings play an important role in diagnosing cardiac and pulmonary diseases. Therefore, information from radiology reports is also significant in diagnosis besides the relationship between symptoms and diseases. This is an area which is left unexplored whereas there are ample studies on just symptom and disease relationships. This paper aims to find clinical findings of internal organs from radiology reports, demographics along with the symptoms and diseases.

From the literature survey [8, 13], it was observed that identifying a patient's medical conditions inclusively from a clinical narrative highly depends on semantical usage of medicinal vocabulary. Biomedical ontologies provide a standard medicinal knowledge base that is built from the contributions of esteemed organizations and universities [22]. This is the reason we have used these ontologies in our paper. The proposed methodology uses five biomedical ontologies for exhaustive analysis and mapping of clinical narratives to KGs, compared to other related works.

This paper presents a framework to curate and enrich KG from clinical narratives. The clinical narratives (radiology reports) contain observations from radiologists/physicians in textual form. Besides, it consists of symptoms, demographics, diseases and radiological findings of patients. Therefore, the narratives have extensive medicinal terminologies in the text. The algorithm, N2K Mapper, in the proposed framework uses NLP, NER, HO and existing biomedical ontologies to enrich and map clinical narratives into KG. N2K

Table 1 Comparative study of mapping of clinical narratives to knowledge structure

Proposed knowledge structure	Methodology	Dataset	Clinical narratives/literature (Input)	Evaluation	
				Method	Criteria
KnowLife [16]	Pattern-based information extraction with confidence and constraint-based consistency checking	PubMed, web portals and online health platforms	Scientific publications, Encyclopedic articles, social sources	Number and precision of facts	Size and quality of KG
TRAK ontology [8]	Ontology, pattern based rules, dictionary lookup, PathNER,	Acute knee screening unit (Cardiff and Vale Health Board)	MRI reports of Knee injuries	Precision, recall, F1-score	Compare manually and automated text mappings of TRAK concepts
KG [10]	String matching, probabilistic models	De-identified patient records	Nursing notes	Precision and recall curve	Against Google's KG and expert physician opinion
Semantic Health KG [17]	First order predicate logic for chain inference, machine learning	Health information system, Zhejiang, China	Chinese notes of doctors (body part, symptom, history, disease, treatment, lab test)	Precision, recall	Chain inference after pruning
TCM KG [18]	Ontology-based database integration (using UMLS)	TCM resources (literature and databases)	Literature, textbooks, papers, dictionaries, references	Not mentioned	–
Biomedical KG [19]	Term extraction: ontology, Naïve Bayes	Biomedical literature	Paper title and abstract	Accuracy	Diagnostic accuracy on symptom checkers
COPD KG [6]	Feature clustering: K-means; feature selection: CMFS-η; classification: DSA-SVM	EMR features	EMR features (smoke, constipation, labour)	Accuracy, recall, and F1-score	Classification of COPD patients
HKGB [5]	Machine learning, long short term memory	The cardiovascular EMR data 2 China hospitals	Unstructured data from EHR (clinical notes)	Precision, recall, human effort	Entity recognition, comparison to save human effort
Quadruplet-based medical KG [14]	BiLSTM, CRF, co-probability	Southwest hospital EMR data	Unstructured data includes Chief complaint and present illness history	Recall, precision, F1-score, normalized DCG	Symptom recognition, ranking symptoms for lung cancer
ASKG [7]	Term extraction: Ontology based IE, skip gram, word embedding, predictive knowledge generation: FOPL and SWRL rules	EHR data from Henry Ford Hospital	Clinical notes (disease, medication, symptom, procedure, risk factor, and aneurysm features)	Precision, recall, and F1-score	Concept and relation extraction, classification, rule evaluation
Health KG [20]	Feature links: Pearson coefficient, latent and domain health knowledge	NHANES	Survey and interview question bank	Accuracy, precision, recall	Disease classification
DSTKG [21]	NER: BiLSTM-CRF event extraction: rule-based SemanticRE: BiGRU-attention network	Chinese EMR, clinical NER challenge of CCKS	Chinese notes mentioning primary diagnosis and medication	Experts' evaluation score, correlation coefficient	Quality evaluation scores based on data, schema and application layer
Biomedical KG [11]	BERT And CRF	MIMIC-III dataset	Clinical narratives (drug, dosage, frequency, Drug adverse effect, drug reason)	F1-score	NER, relation extraction

Table 1 (continued)

Proposed knowledge structure	Methodology	Dataset	Clinical narratives/literature (Input)	Evaluation	
				Method	Criteria
KG [our work]	Rule-based chunking by NLP, NER, Existing Biomedical Ontologies (SYMP, DO, RadLex, anatomy, demographics), and HO	MIMIC-III dataset	Clinical narratives (symptoms, disease, demographics, anatomy (body parts), radiological findings)	Accuracy, precision, recall, F1-score	NER and triple formation

Mapper also enriches existing biomedical ontologies with new medicinal terminologies for future references leading to a sustainable medicinal knowledge base.

The next section details the proposed framework for enriching KG from clinical narratives.

3 Enriching KG for healthcare applications

This section introduces a framework (shown in Fig. 1) to enrich the KG from clinical narratives. The framework consists of N2K Mapper, HO [23] and five pre-existing biomedical ontologies [22, 23]. The patient's history, symptoms, and other relevant findings are extracted from narratives and converted into triples to form a KG.

Many healthcare applications [24] may leverage this N2K Mapper transformed KG to generate recommendations related to patient's health and assist medical practitioners. The N2K Mapper and other components are described below in the subsections.

3.1 Biomedical ontologies

Ontology refers to concepts and their relationships in a domain of interest [25]. In the medicinal domain, one of the practical uses of ontology is to standardize medicinal terminologies that enables standardization during the exchange of medicinal information among different institutes/ organizations [22]. In the proposed framework, these ontologies have been used to identify medicinal terminologies while performing a semantic interpretation of natural language text in the narratives. As shown in the framework, the following ontologies have been used by the N2K Mapper:

3.1.1 Symptom ontology (SYMP)

Change in functioning, appearance, and sensation observed by the patient are known as symptoms. SYMP[1] [26] helps to identify the terms representing symptoms in a patient. SYMP contains more than 900 symptoms. It is maintained by the Institute of Genome Sciences, University of Maryland School of Medicine.

3.1.2 Disease ontology (DO)

It is a standard ontology that contains exhaustive semantic data on diseases that ails human beings. DO was initially developed in 2003 and being updated since then. DO[2] [27] contains names of diseases in a structured hierarchy. It

[1] https://www.ebi.ac.uk/ols/ontologies/symp

[2] https://disease-ontology.org/.

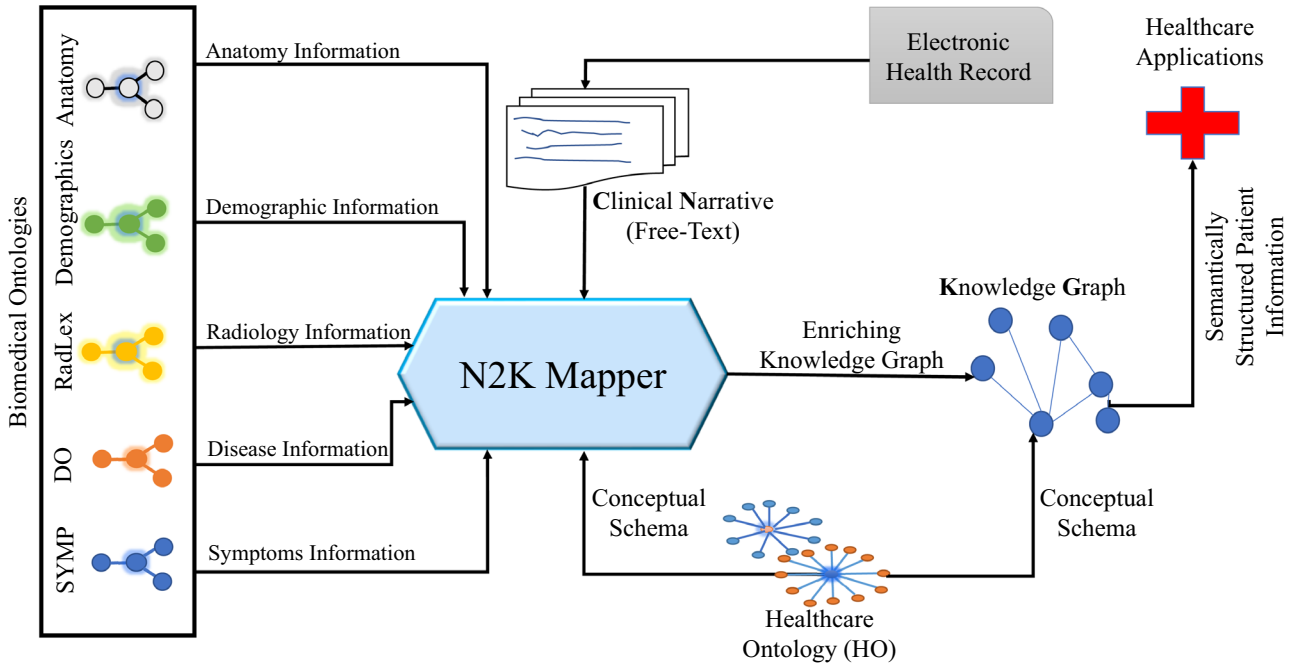

Fig. 1 Framework for enriching KG from clinical narratives using ontologies

contains more than 9000 disease terms, including acquired, inherited and developmental human diseases.

3.1.3 RadLex

It aids in identifying image findings present in a clinical narrative. RadLex[3] [28] is a comprehensive terminology system for medical imaging terms. More than 90 radiologists developed it to annotate musculoskeletal, neurologic, thoracic, pediatric, abdominal, cardiovascular terms, etc. It contains more than 8000 pathological and anatomic terms. It also contains terms describing procedures, devices and imaging techniques, the diagnostic quality of images, etc. RadLex is maintained by the Radiology Society of North America (RSNA).

3.1.4 Anatomy

Foundational model of anatomical ontology (FMA)[4] [29] is used to identify anatomical entities. We have used a slimmed-down version of FMA, which was suitable for our needs.

3.1.5 Demographics

This ontology is developed to store a patient's demographic details such as age, gender, ethnicity, marital status, insurance, and religion. It is used to identify demographic information present in clinical narratives.

These ontologies are updated regularly by their respective maintenance authorities [22, 26, 28]. Therefore, it makes an exhaustive source of medical knowledge. The following sub-section briefs the design of HO and its significance in enriching the KG.

3.2 HO and KG

The proposed system uses HO to map the extracted entities from clinical narratives into a set of triples to constitute a KG. HO forms a conceptual schema for the KG, which provides semantic relationships between entities of a patient's KG.

3.2.1 HO

The HO [23] is designed to capture patients' information stored in EHR. It is a conceptual schema of the KG. The HO semantically represents patients' details using entities (classes or concepts), relationships, and attributes. These entities include Patient (id), Demographics, LabEvents, Reports, Symptoms, Diseases, and Image Findings. HO is

[3] http://radlex.org/.

[4] https://www.ebi.ac.uk/ols/ontologies/fma.

Table 2 Relationships and Attributes between entities in the conceptual schema of KG

Relationship	Domain (entity)	Range (entity)
:hasDemographics	Patient	Demographics
:hasReports	Patient	Reports
:showsSymptoms	Patient	Symptom
:hasPatientDetails	Demographics	Patient details
:reportedDateTime	Report	Date time record
:hasReportedSymptoms	Report	Symptom
:hasReportedDisease	Report	Disease
:hasImageFinding	Report	Image findings
:ofAnatomicalEntity	Image findings	Anatomical entity
:hasReasonForExamination	Report	Reason for examination
:hasView	Report	String

defined as RDFS (Resource Descriptor Framework Schema), a semantic extension of RDF [30].

3.2.2 KG

The KG has two primary components; a conceptual schema and a set of triples. Conceptual schema connects various entities by establishing their relationship (as defined in HO). N2K Mapper uses HO as the conceptual schema to obtain KG from clinical narratives. HO defines appropriate semantic relations between the entities extracted from clinical narratives. For example, a Patient entity is related to demographics with the relationship, hasDemographics with the domain 'patient' and range 'demographics'. Some of the other relationships in HO are shown in Table 2.

Triples [3] are the actual data instances extracted from narratives and mapped into the form of <subject, predicate, object> using conceptual schema. These are the factual details corresponding to an individual patient. For example; <Patient_21, hasReportedSymptoms, Dyspnea>. In the example, 'Patient_21' is an instance of the 'Patient' entity and related to the Symptom entity 'Dyspnea' with the relationship or predicate 'hasReportedSymptoms'. These triples mapped from a particular patient's narrative collectively constitute a patient KG. The next sub-section discusses the N2K Mapper algorithms which utilize NLP, NER to map clinical narrative into a KG.

3.3 Proposed N2K Mapper

The N2K Mapper extracts triples from clinical narratives to enrich the KG. It consists of three components; (1) Preprocessing of narratives, (2) RDF triple extraction, and (3) Results Validation. Figure 2 shows the components of the N2K Mapper.

3.3.1 Pre-processing of narratives

The first component of the N2K Mapper is pre-processing of narratives. It removes redundant information from narratives. The clinical narrative generally has three sections; generic information about the patient, including the patient's name, age, etc., the second section deals with a chief complaint which details the primary problem reported by the patient, and clinical test reports. The last segment contains clinical findings/observations written in natural language by radiologists/physicians. Algorithm 1 (Preprocessing module) shows the pseudocode for the component, pre-processing of narratives. After pre-processing, the resultant narrative is passed on to the second component, RDF triple extraction module.

3.3.2 RDF triple extraction

Algorithm 2 (RDFTripleExtraction module) shows the pseudocode for this component. It is composed of three sub-modules; NLP (Algorithm 3: Get Noun Chunks from Narratives using NLP), NER (Algorithm 4: NER using Biomedical Ontologies), and Triple Formation.

The NLP sub-module uses Chunking [31] to identify and extract significant phrases from the natural text. In chunking, the sentences are broken into smaller units; tokens. Part-Of-Speech (POS) Tagging [32] is applied to these tokens. It groups the tokens into a sentence with their corresponding part-of-speech tag. Chunking takes POS tagged data as input and extracts phrases using a pre-designed grammar pattern. For N2K Mapper, a customized grammar has been designed to find specific patterns according to the observed patterns in the clinical narratives. It leverages the syntactic structure of sentences to extract important entities [33]. For example, to extract the two consecutive nouns from the narrative, the following pattern is used: {<NN.?> + <NN.?>}. The implementation of the same is explained in Sect. 4.

The NER [34] sub-module assigns extracted phrases to their relevant classes using HO and existing biomedical ontologies. It is a process that helps in mapping the identified entities from text into their relevant classes. In N2K Mapper, NER groups the medicinal entities extracted from chunking into their medical classes or categories by utilizing the pre-existing biomedical ontologies. The extracted entity's list and its associated class are passed to either the Triple Formation module or Result Validation based on the output of NER.

The Triple Formation sub-module organizes the identified entities and their associated medical classes in the form of triple <subject, predicate, object> using HO. Two cases/outputs might arise based on the results produced by the

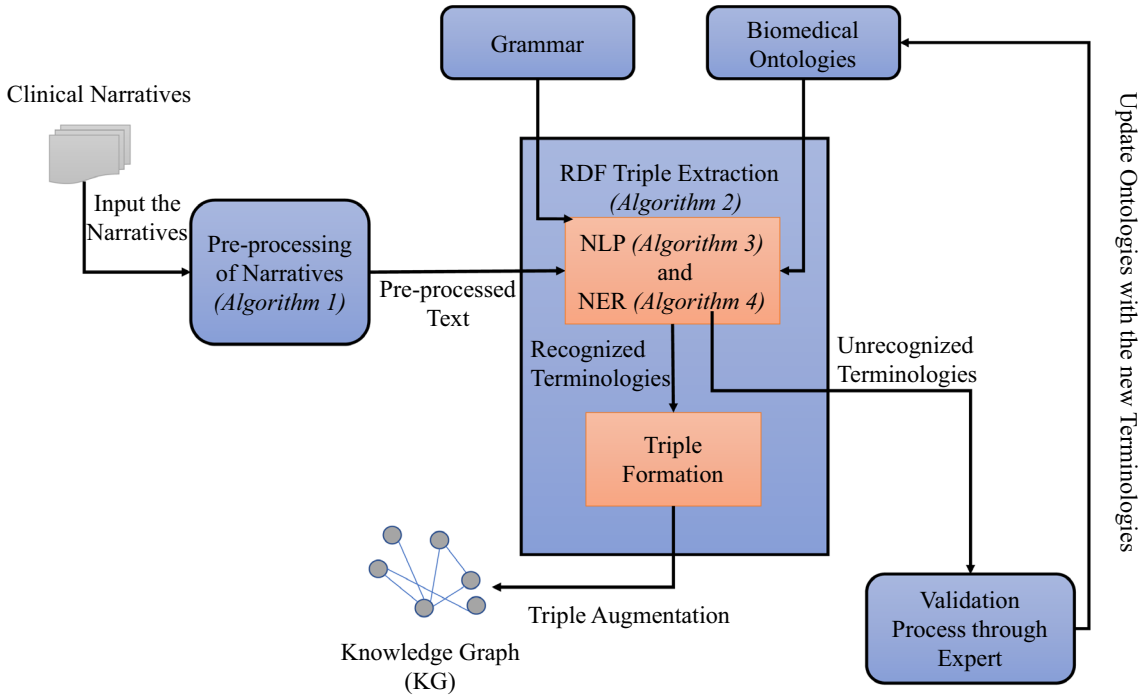

Fig. 2 N2K Mapper

NER function (Fig. 2). These cases and their processing are handled in the third module, 'Result Validation'.

3.3.3 Results validation

The two possible outputs of the NER search of biomedical ontologies: Found and Not Found states the presence and absence of an entity in the ontologies respectively. Based on these findings, the process is divided into the following cases (illustrated in Fig. 3):

Case 1: If an entity extracted from the narrative exists in the biomedical ontologies, it is said to be a 'recognized entity' that follows the Case 1 process. A recognized entity fulfills the condition to form a valid triple. Therefore, it is passed to the Triple Formation function. The valid triple is later augmented to the KG.

Case 2: Alternatively, if the extracted entity does not exist in the biomedical ontologies, it is said to be a 'unrecognized entity' and sent to a domain expert for validation. An expert validates the entity, and the entity is augmented in a local copy of medicinal ontology. It enriches the biomedical ontologies, which help in subsequent extractions.

The next section presents an experimental study using the MIMIC-III dataset [35].

4 Experimental study

An experimental study was conducted on 92 chest X-ray reports of 10 patients taken from the MIMIC-III [35] dataset. MIMIC is a freely available dataset used by researchers for their experiments [11]. The dataset consisting of chest X-ray reports was extracted from the notes table of the MIMIC-III dataset using Google's BigQuery [36] platform. All the algorithms were coded using Python programming language and Jupyter Notebook as Integrated Development Environment (IDE). Following python libraries were used in the experiment: NLTK [37], Owlready2 [38], SPARQL [39], spacy, re, openxyl, pandas, and json.

The chest X-ray reports were mapped to KG using the N2K Mapper. Processing of these reports with exemplar instances using the different components of the N2K Mapper is discussed below.

4.1 Pre-processing

In pre-processing, redundant information in narratives including *"comma (,), next line (\n), hyphen (-), multiple spaces, hash (#), asterisk (*), inverted commas ("")"* were removed. The narratives were divided into three segments in the pre-processing step as explained in Sect. 3.3 and were fed to the RDF Triple Extraction module.

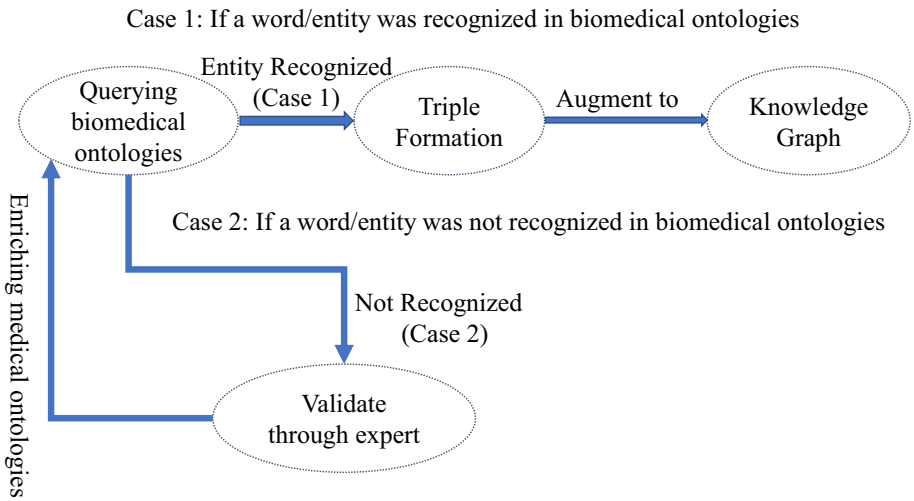

Fig. 3 Cases obtained based on the results (found/not found) while searching in biomedical ontologies

4.2 RDF triple extraction

The chest X-ray narratives were parsed into sentences through a RegexParser of the NLTK library. These sentences are tagged using POS tagging. The POS tags were referred from Penn Tree Bank. RegexParser contains grammatical patterns which return matched chunks or entities from the parsed chest X-ray. These chunks were passed onto the NER.

N2K Mapper (main module)

```
procedure accessNarratives(dataset x):
    for i = 1 → len(x) do
        nar = readFromDataset(i)
        pp = preprocessing(nar)
        return RDFTripleExtraction(pp)
    end for
end procedure
```

Algorithm 1. Preprocessing module

```
procedure preProcessing(nar):
    text = removeExtraSymbols(nar)
    suparts = splitSegments(text)
    return subparts
end procedure
```

Algorithm 2. RDFTripleExtraction Module

```
procedure RDFTripleExtraction (text):
    nc = getNounChunks(text)
    for i in nc then
        rel = searchBiomedicalOntology(i)
        if found then
            tripleFormation(i, rel)
        end if
        else
            resultValidation(i)
        end else
    end for
end procedure
```

Algorithm 3. Get Noun Chunks from Narratives using NLP

```
procedure getNounChunksList(pp):
    NounChunksList = []
    sent = sent_tokenize(pp)
    chunker = getGrammar()
    for i in sent do
        nc = createNounChunks(i, chunker)
        nc = exceptionHandling(nc)
            for y in nc do
                if y is compoundWord then
                    n_nc = handleCompoundWord(y)
                    appendToNounChunksList(n_nc)
                end if
                else
                    appendToNounChunksList(y)
                end else
    return NounChunksList
```

Algorithm 4. NER using Biomedical Ontologies

```
procedure NERUsingBiomedicalOntology (NounChunk):
findingWord = NounChunk
setOntologyPrefix()
on_name = queryOntology(findingWord)
if found then:
    rel = setRelationship(on_name)
    return rel
end if
else
    return not found
end else
end procedure
```

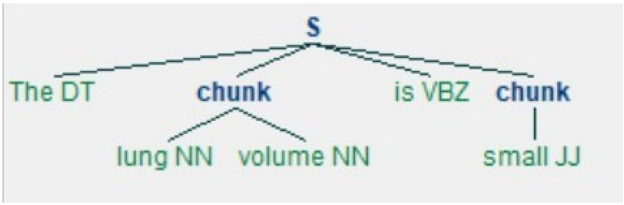

Fig. 4 Chunks extracted in RDF Triple Extraction module using NLTK

Consider a sentence in Narrative: *The lung Volume is small.* The sentence was parsed using the RDF Triple Extraction module, where each word in the sentence was assigned a POS tag, such as *the* = Determinant (DT), *lung* = common Noun (NN), *volume* = common Noun (NN), *is* = Verb (VBZ), *small* = Adjective (JJ). Thus, the parsed sentence was converted to a tree structure. The compound words or chunks were extracted from the resulting tree and passed to the entity recognition algorithm (NER). A visual representation of one of the extracted chunks using the NLTK library is shown in Fig. 4.

The NER algorithm (Algorithm 4) used SPARQL to search the entities in SYMP, DO, RadLex, Anatomy and Demographics ontologies to determine an appropriate category of the entity. Owlready2, a python library was used as an 'end point' for SPARQL. The entities along with their assigned categories were passed to Result Validation component as input.

4.3 Result validation

After executing *Case 1* (recognized entity) and *Case 2* (unrecognized entity), results were validated by a domain expert for evaluation purposes. Individual instances of both the cases are demonstrated below.

Case 1: A word, 'dyspnea' was identified in clinical narratives. It was found in SYMP ontology. The triple formed was <21><hasReportedSymptom><dyspnea>. It was then augmented to the KG.

Case 2: A word identified in clinical narratives was 'lung ca'. It was not found in any of the ontologies. After an expert's opinion, it was concluded that the word was referring to lung cancer disease. It was updated in the local copy of DO for future references.

After the augmentation, *Case 1* was followed to complete the triple formation process, which resulted in the triple: <21><hasReportedDisease><Lung cancer> as a part of KG.

The process of enriching the KG from a sample Chest X-ray report is shown diagrammatically in Fig. 5.

The next section shows the results obtained during the experimental study conducted with the MIMIC-III dataset.

5 Results

The N2K Mapper algorithm was able to identify 1630 unique entities from the dataset of clinical narratives. It generated a KG consisting of 4174 RDF triples using the identified entities and semantic information retrieved from the clinical narratives. Out of the 4174 triples, 3347 references were found in biomedical ontologies (*Case 1*), whereas 827 references were augmented in biomedical ontologies with the help of an expert's feedback (*Case 2*). It shows that more than 80% of the references entities from clinical narratives were recognized using biomedical ontologies. It proves that the existing biomedical ontologies taken together are a good resource of medicinal information. The performance of the N2K Mapper was evaluated using accuracy, precision, recall and F1-score [40] metrics.

A confusion matrix [40] for the N2K Mapper is shown in Fig. 6. The matrix consists of the number of actual triples and the triples mapped by the N2K Mapper over the four labels: disease, symptom, image findings, and others. It shows 81.5% accuracy, which means that 81.5% of the total triples have been correctly mapped onto their corresponding labels.

Precision and recall for the N2K Mapper (calculated as mean of precision and recall of each group of the labels) were evaluated to be 78.54% and 85.07% respectively. F1-score, which is the harmonic mean of precision and recall, was evaluated to be 78.97%. The results of the proposed work have been compared with similar published work on KG mapping from unstructured clinical narratives. Although the compared approaches are implemented on different datasets (due to different sets of clinical notes and selected features taken as input), they determine the validity of triples in terms of precision and recall. Table 3 shows the comparison of results obtained by the N2K Mapper with those published in similar works. Though the N2K Mapper shows slightly better precision, it has a far better recall measure (85.07%) in comparison to other related work (71% of [7] and 71.31% of [8]). In the medicinal domain, recall is more significant because it reduces the misclassification of positive cases as negatives. In FlexiTerm [8], the stemming of tokens to normalize the entities and using the bag of word technique in term hood calculation increases the precision value. However, utilizing these techniques loses the semantics of entities and can increase the false negatives. Thus, the N2K Mapper shows overall better precision and recall over similar published work.

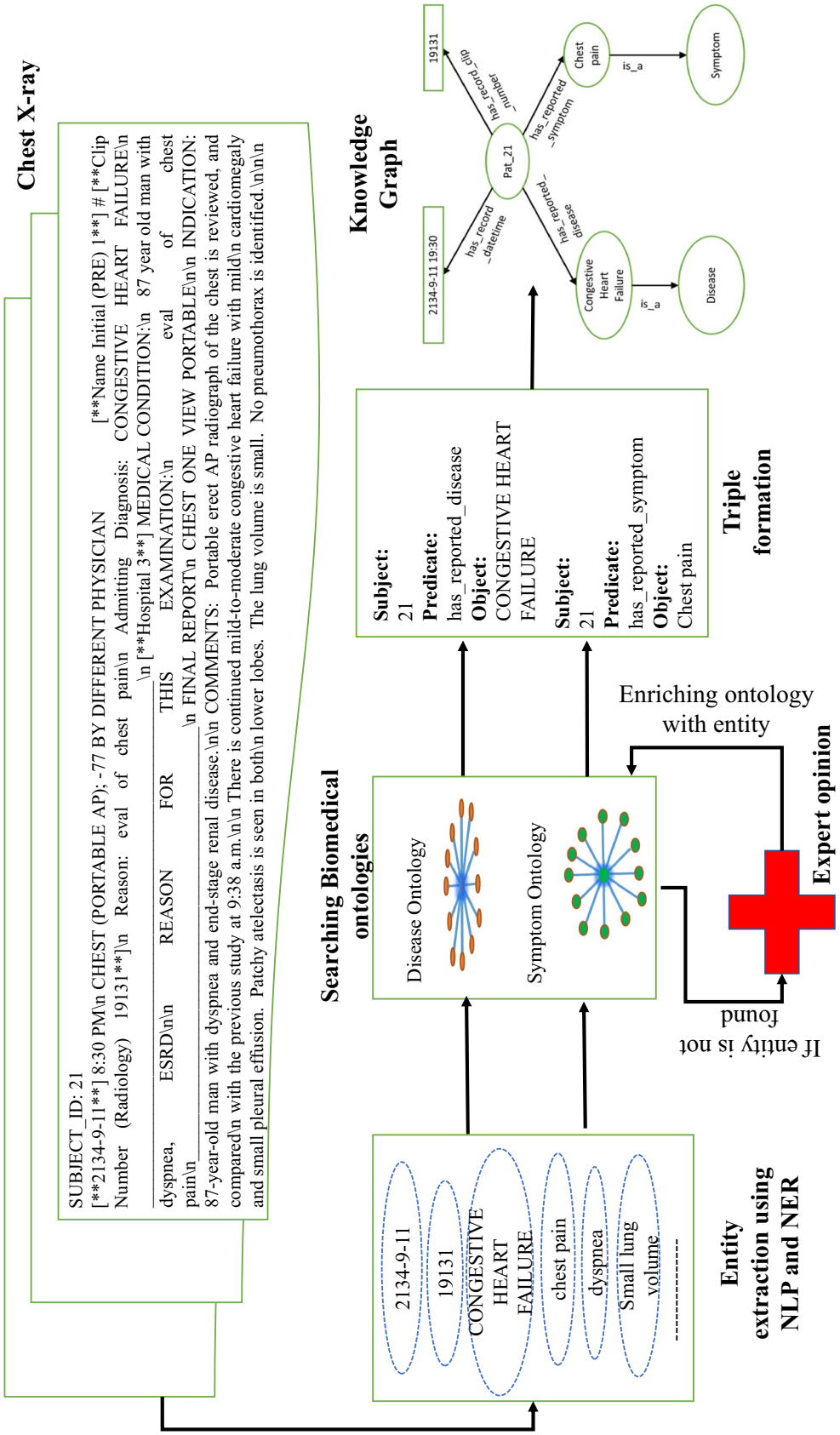

Fig. 5 Enriching KG from a chest X-ray sample

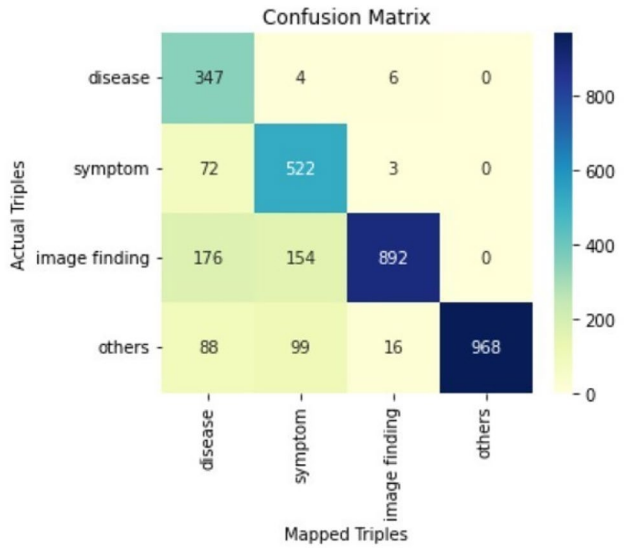

Fig. 6 Confusion matrix representing actual triples versus mapped triples of four labels

Table 3 Comparison of the N2K Mapper with similar work

Approach/method[a]	Precision (%)	Recall
KnowLife [16]	78.4	Not Mentioned
(OBIE + Word Embedding) [7]	78	71%
FlexiTerm [8]	94.56	71.31%
N2K Mapper (Our Work)	78.54	85.07%

[a]These approaches are implemented on different datasets, however they determine validity of triples in terms of precision and recall

6 Conclusion and future scope

The N2K Mapper consisting of a set of algorithms was presented in the paper. The clinical narratives containing clinical evidence exist in natural text. As a result, AI-based healthcare applications are not able to directly extract or semantically process patients' clinical evidence from the narratives. N2K Mapper was designed to analyze the narratives and map those into a semantically structured format called KG. The Mapper firstly pre-processed the data where data cleaning was done. Secondly, the pre-processed narratives were passed to the RDF triple extraction module, consisting of two sub-modules, NLP and NER. They are responsible for extracting chunks or entities from the clinical narratives and assigning them with their relevant classes with the help of existing biomedical ontologies. Another module in RDF triple extraction was 'triple formation'. It semantically mapped the extracted entities into <subject, predicate, object> triples. The entities which were not found in the ontologies were verified manually by an expert and augmented to the ontologies. This process helped in enriching the biomedical ontologies and subsequently the KG. An experimental study was conducted on 92 chest X-ray reports extracted from the MIMIC-III dataset. The N2K Mapper algorithm successfully mapped 4174 triples from the 92 clinical narratives. Out of the total mapped triples, 3347 references were found in the biomedical ontologies, while 827 were augmented in ontologies by an expert's feedback. Evaluation of the N2K Mapper showed substantially good performance with an accuracy of 81.5%, precision and recall of 78.5% and 85.07%, respectively, and an F1-score of 78.97%. In future, we intend to create a healthcare application for differential disease diagnosis based on the proposed framework and evaluate the performance with real-time data.

Data Availability The data that support the findings of this study are available from [35] (https://physionet.org/content/mimiciii/1.4/), but restrictions apply to the availability of these data, which were used under license for the current study.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Nurdiati S, Hoede C (2008) 25 years development of knowledge graph theory: the results and the challenge. Memorandum 1876(2):1–10
2. Singhal A (2012) Introducing the knowledge graph: things, not strings. Official Google Blog
3. Ji S, Pan S, Cambria E, Marttinen P, Yu PS (2021) A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2021.3070843
4. Tian L, Zhou X, Wu YP, Zhou WT, Zhang JH, Zhang TS (2022) Knowledge graph and knowledge reasoning: a systematic review. Appl Geochem. https://doi.org/10.1016/j.jnlest.2022.100159
5. Zhang Y et al (2020) HKGB: an inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. Inf Process Manag 57(6):102324. https://doi.org/10.1016/j.ipm.2020.102324
6. Fang Y, Wang H, Wang L, Di R, Song Y (2019) Diagnosis of COPD based on a knowledge graph and integrated model. IEEE Access 7:46004–46013. https://doi.org/10.1109/ACCESS.2019.2909069
7. Malik KM, Krishnamurthy M, Alobaidi M, Hussain M, Alam F, Malik G (2020) Automated domain-specific healthcare knowledge graph curation framework: subarachnoid hemorrhage as phenotype. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2019.113120
8. Spasić I, Zhao B, Jones CB, Button K (2015) KneeTex: an ontology-driven system for information extraction from MRI reports. J Biomed Semantics 6(1):1–26. https://doi.org/10.1186/s13326-015-0033-1
9. Yao L, Liu H, Liu Y, Li X, Anwar MW (2015) Biomedical named entity recognition based on deep neutral network. Int J Hybrid Inf Technol 8(8):279–288. https://doi.org/10.14257/ijhit.2015.8.8.29

10. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D (2017) Learning a health knowledge graph from electronic medical records. Sci Rep 7(1):1–11. https://doi.org/10.1038/s41598-017-05778-z
11. Harnoune A, Rhanoui M, Mikram M, Yousfi S, Elkaimbillah Z, El Asri B (2021) BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Comput Methods Progr Biomed Update 1:100042. https://doi.org/10.1016/j.cmpbup.2021.100042
12. Nath N, Lee SH, McDonnell M, Lee I (2021) The quest for better clinical word vectors: ontology based and lexical vector augmentation versus clinical contextual embeddings. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2021.104433
13. Kamdar MR et al (2020) Text snippets to corroborate medical relations: an unsupervised approach using a knowledge graph and embeddings. AMIA Summits Transl Sci Proc 2020:288–297
14. Li L et al (2020) Real-world data medical knowledge graph: construction and applications. Artif Intell Med 103:101817. https://doi.org/10.1016/j.artmed.2020.101817
15. Yuan H, Deng W (2021) Doctor recommendation on healthcare consultation platforms: an integrated framework of knowledge graph and deep learning. Internet Res. https://doi.org/10.1108/INTR-07-2020-0379
16. Ernst P, Siu A, Weikum G (2015) "KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics. https://doi.org/10.1186/s12859-015-0549-5
17. Shi L, Li S, Yang X, Qi J, Pan G, Zhou B (2017) Semantic health knowledge graph: semantic integration of heterogeneous medical knowledge and services. Biomed Res Int. https://doi.org/10.1155/2017/2858423
18. Yu T et al (2017) Knowledge graph for TCM health preservation: design, construction, and applications. Artif Intell Med 77:48–52. https://doi.org/10.1016/j.artmed.2017.04.001
19. Xia E, Sun W, Mei J, Xu E, Wang K, Qin Y (2018) Mining disease-symptom relation from massive biomedical literature and its application in severe disease diagnosis. AMIA Annu Symp Proc 2018:1118–1126
20. Tao X et al (2020) Mining health knowledge graph for health risk prediction. World Wide Web. https://doi.org/10.1007/s11280-020-00810-1
21. Xiu X, Qian Q, Wu S (2020) Construction of a digestive system tumor knowledge graph based on Chinese electronic medical records: development and usability study. JMIR Med Inform. https://doi.org/10.2196/18287
22. Smith B et al (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11):1251–1255. https://doi.org/10.1038/nbt1346
23. Dhiman S, Thukral A, Bedi P (2022) OHF: an ontology based framework for healthcare. In: Dev A, Agrawal SS, Sharma A (eds) Artificial intelligence and speech technology: AIST 2021. Communications in computer and information science, vol 1546. Springer, Cham, pp 318–328
24. Vidhate DA, Kulkarni P (2018) Improved decision making in multiagent system for diagnostic application using cooperative learning algorithms. Int J Inf Technol. https://doi.org/10.1007/s41870-017-0079-7
25. Gruber T (2009) Definition of ontology. Database systems, pp 10–12
26. Mohammed O, Benlamri R, Fong S (2012) Building a diseases symptoms ontology for medical diagnosis: an integrative approach. In: 1st international conference on future generation communication technologies, FGCT 2012, pp 104–108. https://doi.org/10.1109/FGCT.2012.6476567
27. Schriml LM et al (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 47(D1):955–962. https://doi.org/10.1093/nar/gky1032
28. Langlotz CP (2006) RadLex: a new method for indexing online educational materials. Radiographics. https://doi.org/10.1148/rg.266065168
29. Rosse C, Mejino JLV (2003) A reference ontology for biomedical informatics: the foundational model of anatomy. J Biomed Inform 36(6):478–500. https://doi.org/10.1016/j.jbi.2003.11.007
30. Cyganiak R, Wood D, Lanthaler M (2014) RDF 1.1 concepts and abstract syntax. W3C Recommendation
31. Tjong Kim Sang EF, Buchholz S (2000) Introduction to the CoNLL-2000 shared task: Chunking
32. Akhil KK, Rajimol R, Anoop VS (2020) Parts-of-Speech tagging for Malayalam using deep learning techniques. Int J Inf Technol. https://doi.org/10.1007/s41870-020-00491-z
33. Thanawala P, Pareek J (2018) MwTExt: automatic extraction of multi-word terms to generate compound concepts within ontology. Int J Inf Technol. https://doi.org/10.1007/s41870-018-0111-6
34. Sintayehu H, Lehal GS (2021) Named entity recognition: a semi-supervised learning approach. Int J Inf Technol. https://doi.org/10.1007/s41870-020-00470-4
35. Johnson AEW et al (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3(1):1–9. https://doi.org/10.1038/sdata.2016.35
36. Sato K (2012) An inside look at Google BigQuery. Google Inc
37. Loper E, Bird S (2002) NLTK: the natural language Toolkit. Assoc. Comput. Linguist.
38. Lamy JB (2017) Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artif Intell Med 80(2020):11–28. https://doi.org/10.1016/j.artmed.2017.07.002
39. DuCharme B (2010) Learning SPARQL querying and updating with SPARQL 1.1
40. Zaki MJ, Meira W Jr (2014) Data mining and analysis: fundamental concepts and algorithms, 2nd edn. Cambridge University Press, Cambridge