Journal of Biomedical Informatics 98 (2019) 103276

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Ontology-based clinical information extraction from physician's free-text notes

Engy Yehia[a,b,*], Hussein Boshnak[c], Sayed AbdelGaber[a], Amany Abdo[a], Doaa S. Elzanfaly[a]

[a] Information Systems Department, Faculty of Computers and Information, Helwan University, Helwan, Cairo, Egypt
[b] Business Information Systems Department, Faculty of Commerce and Business Administration, Helwan University, Helwan, Cairo, Egypt
[c] General Surgery Department, Faculty of Medicine, Ain Shams University, Cairo, Egypt

ARTICLE INFO

Keywords:
Information extraction
Electronic health records
Natural language processing

ABSTRACT

Documenting clinical notes in electronic health records might affect physician's workflow. In this paper, an Ontology-based clinical information extraction system, OB-CIE, has been developed. OB-CIE system provides a method for extracting clinical concepts from physician's free-text notes and converts the unstructured clinical notes to structured information to be accessed in electronic health records. OB-CIE system can help physicians to document visit notes without changing their workflow. For recognizing named entities of clinical concepts, ontology concepts have been used to construct a dictionary of semantic categories, then, exact dictionary matching method has been used to match noun phrases to their semantic categories. A rule-based approach has been used to classify clinical sentences to their predefined categories. The system evaluation results have achieved an F-measure of 94.90% and 97.80% for concepts classification and sentences classification, respectively. The results have showed that OB-CIE system performed well on extracting clinical concepts compared with data mining techniques. The system can be used in another field by adapting its ontology and extraction rule set.

1. Introduction

Electronic Health Records (EHRs) are computerized information systems that used to collect, store and display patient information [1]. The development of electronic health records has necessitated the use of innovative technologies to support the transition from paper-based records for healthcare providers [61]. Despite the high expectations and interest in EHRs worldwide, the rate of EHR acceptance by physicians remains slow in some countries and they face several problems. For instance, EHRs adoption require significant financial investment, they are seen as contrary to a physician's traditional working style, and they require a greater capability in dealing with computers and installing a system [11,26]. In this work, we focus on the extraction of clinical concepts from clinical documents which written by a physician using his/her pen through the patient visit. We use a rule-based method which requires less resources and is easier to adapt to a new domain rather than using annotated corpora to train learning algorithms. The rules we used in our IE system is based on a clinical domain ontology. In knowledge representation, an ontology is a description of the concepts and relationships in an application domain [3]. The main reason behind using a domain ontology in our information extraction system is that the rules that are based on ontological concepts used in entity recognition and information extraction, have more expressive power than those which are based on textual items [58]. The main objective of this work is to propose an Information Extraction System that allow for better documentation that facilities the physician work and assist in recognizing unstructured information in EHRs. The system also can be implemented on other domain by changing the ontology knowledge source.

2. Related work

There were several studies conducted in community physician offices, mixed healthcare settings and hospitals which examined EHR implementations in hospitals and their impact on physician work. These studies reported that the people barriers are the most important perceived barriers among EHR barriers [7,8,10,11,12,15,19,23,24,26,27,30,46]. According to Carayon, P., et al study results, EHR

* Corresponding author at: Information Systems Department, Faculty of Computers and Information, Helwan University, Helwan, Cairo, Egypt. Business Information Systems Department, Faculty of Commerce and Business Administration, Helwan University, Helwan, Cairo, Egypt. 45 A Thabet Street, Helwan, Cairo, Egypt.
E-mail address: engy_yehia@commerce.helwan.edu.eg (E. Yehia).

https://doi.org/10.1016/j.jbi.2019.103276
Received 26 April 2019; Received in revised form 17 August 2019; Accepted 28 August 2019
Available online 29 August 2019

technology has a major impact on physician work indicating an increase in amount of time spent on documentation [15]. Some Physicians reported that Using EHRs will take more time for each patient than using paper as, in some situations; it might be more convenient for physicians to use paper records during the patient session [8,11]. According to standard procedures, all office visits should be documented by an accompanying visit note. The notes module of the EHR allows for multiple types of data entry including dictation and typed entry, both of which produce unstructured free form text notes [47]. Free text is a more expressive and natural method to document clinical proceedings and facilitate communication among the care team in the health care institutions. The information extraction tasks can use to integrate the free text notes in the EHR data for clinical decision support, quality improvement, or clinical research, by automatically extracts and encodes clinical information from text [63]. Information Extraction IE is an area of Natural Language Processing (NLP) that deals with finding structured objects in free text, such as database records. IE identifies a predefined set of concepts in a specific domain, where a domain consists of a corpus of texts together with a clearly specified information need [41,45]. NLP techniques are used to analyze the text before extracting information from it. The analysis is done using a set of fundamental techniques such as tokenization, sentence detection, classification, and extracting relationships. NLP addresses areas of IE such as extracting predefined types of information from text, speech processing, summarization of text, relationship extraction, and document categorization [17,48,63].

In this paper, we introduce an information extraction system that convert the physician's unstructured clinical notes to structured information to be stored and accessed in EHR system without affecting or changing the physician's time and workflow. There are two basic approaches for clinical information extraction: rule-based and machine learning. In the rule-based approach, a human- knowledge engineer identifies the required knowledge and the extraction rules of the IE system component and mainly uses dictionary lookup and rules. Machine learning–based IE approaches focuses on producing training data, corpus statistics or rules are then derived automatically from the training data and used in analyzing novel texts [4,21,63]. Several recent researches in biomedical informatics has centered on recognizing named entity of various medical entities in clinical text. While most of the Named Entity Recognition (NER) methods are rule-based [5,25,40,50], other systems apply hybrid approaches combing machine learning and rules [18,20,28,51]. According to Y. Wang et al. study, the most frequently used tools for IE in the clinical domain are cTAKES that applies hybrid approaches, MetaMap, and MedLEE which are examples of a rule-based systems [63]. cTAKES is an open-source Natural Language Processing system that combines rule-based and Machine Learning techniques for information extraction from free texts in electronic medical record. It provides a comprehensive platform for performing many clinical information extraction tasks such as syntactic and semantic parsing [51]. YTEX is a series of extension modules on top of cTAKES that provides a generalizable framework for mapping clinical phrases from any domain ontology to various terminologies [28]. MedTAS/P is an extensible and modifiable knowledge representation model which uses natural language processing principles, machine learning and rules to automatically extracting cancer disease characteristics from free-text pathology reports [18]. MedLEE is a tool for processing clinical text that is used for vocabulary development and encoding. MedLEE was originally developed for the domain of radiological reports of the chest, but has subsequently been extended to multiple domains and applications [25]. MetaMap is developed to map scholarly biomedical text to the UMLS Metathesaurus by the National Library of Medicine (NLM). MetaMap uses a knowledge-intensive approach based on symbolic, Natural Language Processing and Computational Linguistic techniques [5,6]. DNorm is an example of machine-learning approach that is based on pairwise learning for ranking, normalizing disease names in biomedical text and computing similarities between mentions and concept names directly from the training data [37]. In the rule-based IE systems, the rule could be developed through manual knowledge engineering and leveraging knowledge base, or a hybrid system [63].

Some of the prior studies attempted to solve the changing in physician workflow barriers by let the physician writes the patient visit notes as unstructured free text and then converts these unstructured notes to structured data [22,29,33,35,52,54,55,57,60,64]. According to These studies the physician doesn't have to enter the patient's notes using forms of data entry such as: drop-down menus, check boxes, prefilled templates and radial buttons, which will be led to save the physician's time and interact efficiently with the patient. However, the physicians still have to enter this free text throw graphical user interface using keyboard typing which may lead to loss the interaction with the patient and it may be time consuming for the physician. The advantage of our method over these methods in biomedical named entity extraction is that our information extraction system doesn't only extract information from clinical notes, but also store the extracted information in the structured EHR database without additional effort or time from physicians which solve the problem of physician's resistance to implement EHR system. Current methods used to extract clinical information from a free machine-readable text written by a physician, while our proposed method extract clinical information from a physician's handwritten notes which leads to save physician time to interact with his/her patient instead of interacting with the computer screen. Another advantage of our proposed system is the proper usability of the extracted information which consider one of the main problems of the current IE systems in medical domain. The Patient Clinical Data (PCD) ontology, the information model for the proposed system, created based on clinician's and domain expert's opinions and evaluated using real patient clinical notes. So, the proposed system can be applied on real medical practices for enhancing physician's work and recognizing unstructured data in the EHR system which can be used in decision making and knowledge discovery. We believe that using PCD ontology for information extraction is important in the sense that it does not only represent concepts with their semantic groups for clinical knowledge domain, but also has a structure for patient clinical data stored in the EHR system. The PCD ontology represents concepts as classes and individuals, each class has a relationship with other classes (object property) and has a list of data properties which built according to HL7 standard. Moreover, the mapping of this ontology with the EHR database ensures a step forward to convert the unstructured clinical notes to structured data to be accessible in the EHR database, and thus, can further be reasoned by decision support systems. In addition to integrating UMLS knowledge sources while developing PCD ontology, we investigated real patient data collected from different medical specialists, written by physicians and stored in the EHR database to ensure that the PCD ontology concepts and structure cover all the patient data produced through the healthcare activities and stored in the EHR system.

3. Materials and methods

An Ontology-based Clinical information extraction system (OB-CIE) is proposed to extract the clinical information from free-text clinical notes and convert them into a structured information. We integrated a domain ontology called patient clinical data (PCD) to be used as a domain knowledge in the (OB-CIE) system [13]. The physician's handwritten clinical notes document is the input of the OB-CIE system which scanned to the computer and processed through several modules as shown in the illustrate example in Fig. 1. The output of the OB-CIE system is a structured information to be recognized in the EHR database.

OB-CIE system consists of three main modules: the data pre-processing module, the Natural Language Processing module and the Information Extraction module. Fig. 2 shows the architecture of the OB-CIE system.

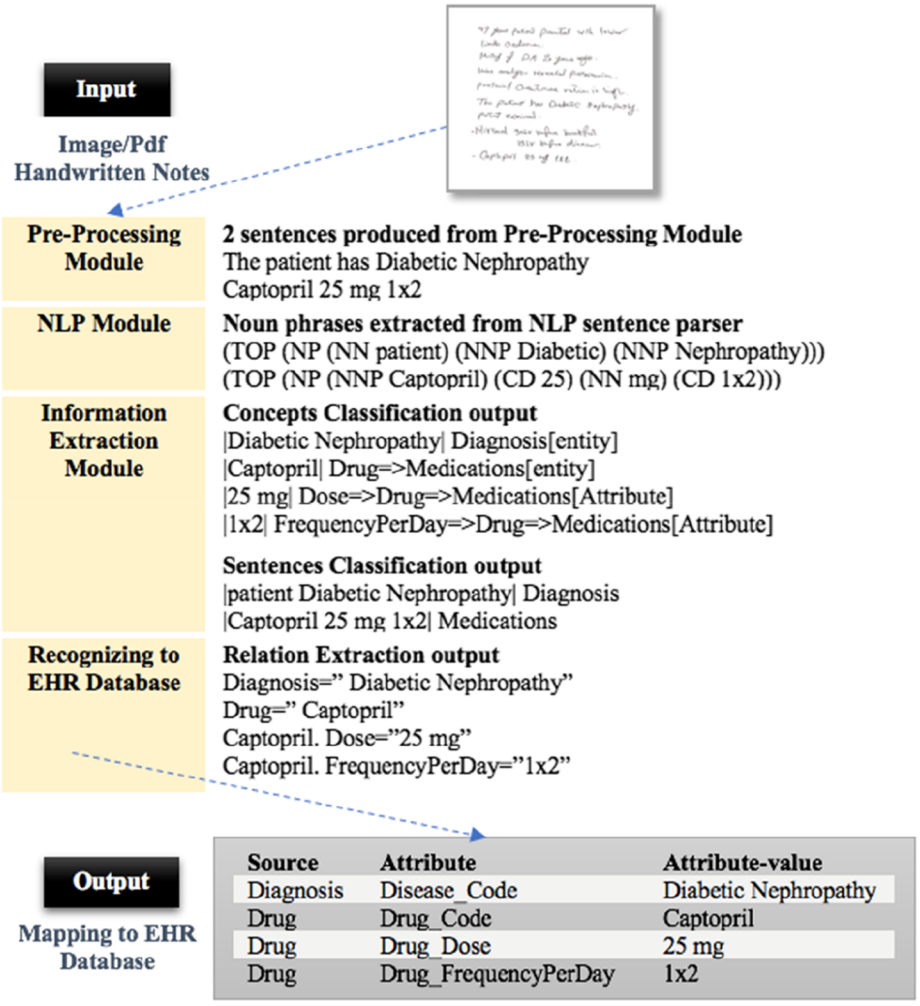

Fig. 1. An overview of the OB-CIE system with an illustrate example.

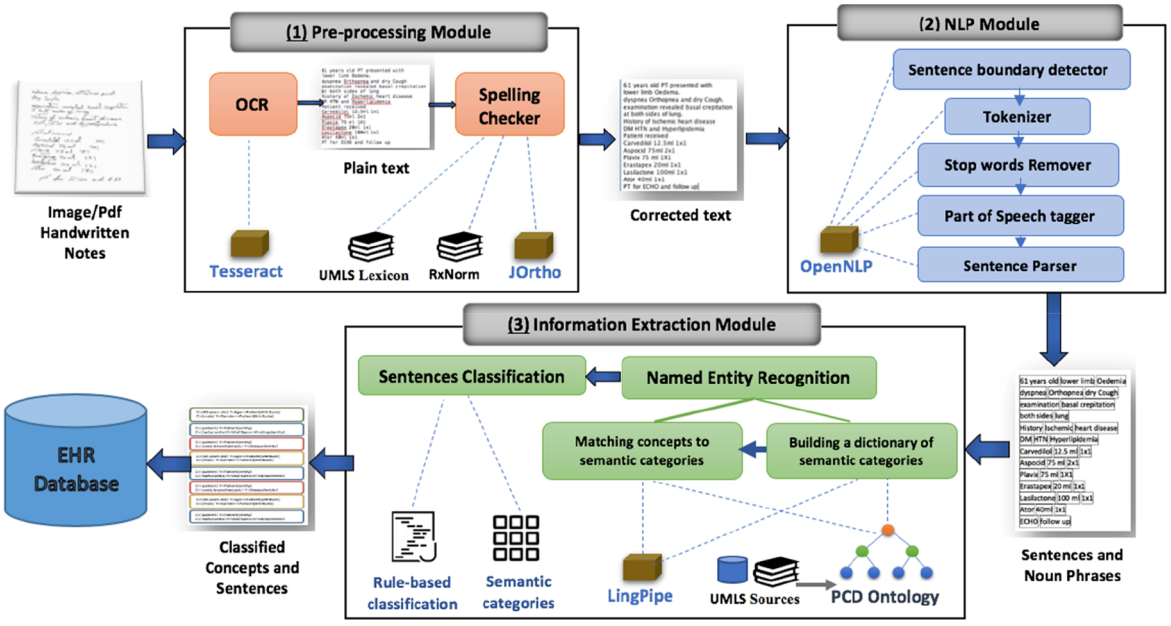

Fig. 2. The OB-CIE system architecture.

3.1. Data pre-processing module

In data pre-processing module, the original handwriting clinical notes are converted to text to be ready for processing in the natural language processing module. To pre-process handwriting clinical notes, the clinical notes passes through two phases including Optical Character Recognition and Spell Checking.

3.1.1. Optical character recognition (OCR)

Optical Character Recognition (OCR) is the process of converting scanned images of handwritten or typewritten text into machine editable text documents through several stages of a computer recognition system [16]. Handwriting recognition is the ability of a computer to receive and interpret handwritten input from sources such as paper documents, touch- screens and other devices. Handwritten documents can have various font types or writing styles. The writing styles can be categorized into discrete style (handprint or boxed style), continuous style (cursive style), and mixed style [16].

In the OB-CIE system, the physician uses his/her pen to write patient visit notes in a paper. Then this paper is scanned to the computer and saved as image. OCR component is used to recognize the handwriting text and convert into editable text file. Tesseract OCR engine is used to build the OCR component. Tesseract is an open-source OCR engine that was developed at HP between 1984 and 1994 and is now one of the most reliable open-source OCR Engine in terms of its accuracy [56]. Tess4J is used to integrate Tesseract OCR engine in the OB-CIE system. Tess4J is a Java JNA wrapper for Tesseract OCR API that released and distributed under the Apache License, v2.0 [59].

3.1.2. Spell checking

The output text from the OCR phase may produce misspellings in the recognized text specially if the scanned handwritten documents have various writing styles. Spell checking is an important phase to check the correctness of each concept before moving to the NLP module. OB-CIE spell checker processes the text in two steps: detects the potentially wrong-spelled concepts, then converts them to the correct form. JOrtho (Java Orthography) is used to build the spell checker component in the OB-CIE system. JOrtho is an open source library entirely written in Java and its dictionary is based on the free Wiktionary project that contains 5,836,006 entries with English definitions from over 3800 languages [34]. To build the spell checker of the OB-CIE system, JOrtho dictionary has been customized by integrating UMLS SPECIALIST Lexicon and RxNorm concepts [49,62]. The process of spell checking is shown in Fig. 3. Firstly, the spell checker checks the correctness of each term by looking up it in the dictionary. If the term is not found, the spell checker detects and highlights it as a misspelled term. Secondly, the checker looks up in the dictionary to generate a list of correct term suggestions and ranking them to let the user choose the appropriate one. Finally, If the term is valid but wasn't found in the dictionary, the user can add it to the dictionary. For example, if we have the following sentence:

{61 years old PT presented with lower limb Oedena, dyspnea, Orthopnea and dry Cough}

The spell checker detects and highlights "Oedena" and "Orthopnea" as a wrongly spelled term, then produces a context menu of a possible correction forms of those terms. After that, the user checks the validation of each term. If the term is not valid such as "Oedena", the user checks the correct form from a context menu such as {Oedema, Medean, edema, Medina}and replaces "Oedena" with "Oedema". If the term is valid such as "Orthopnea", this mean that "Orthopnea", is not found in the dictionary, so the user adds it to the dictionary.

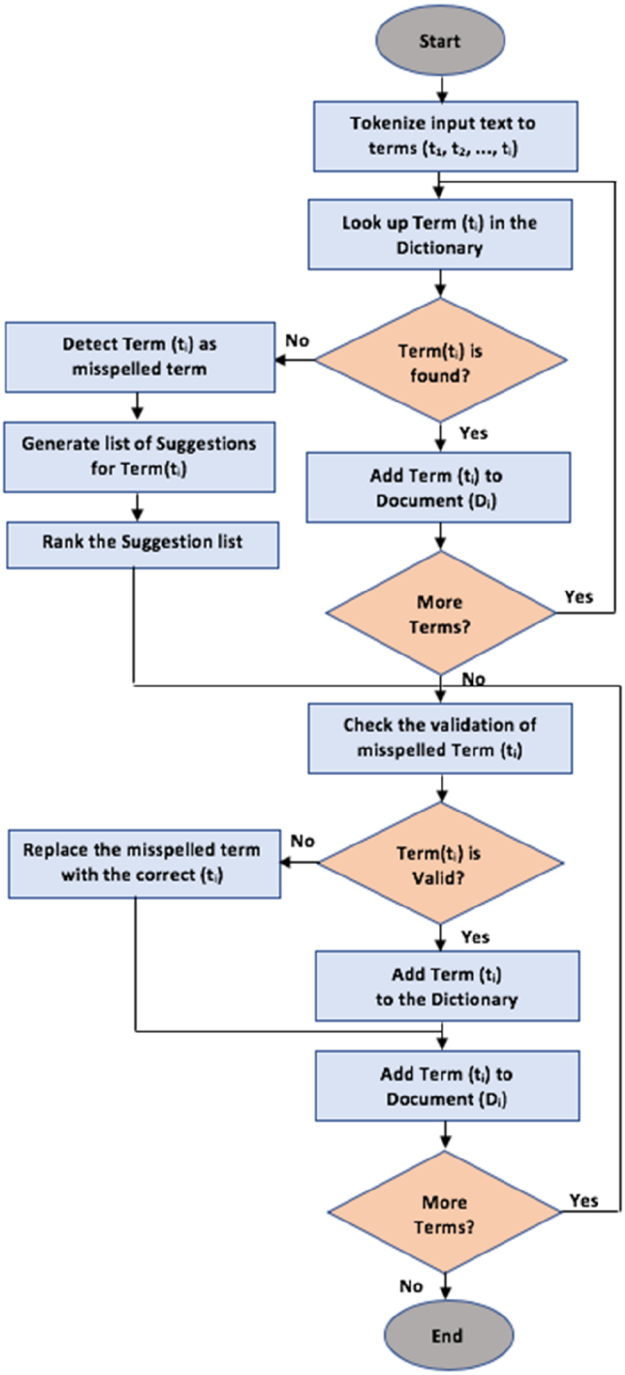

Fig. 3. The flowchart of spell-checking process.

3.2. Natural language processing module

Information Extraction typically requires some pre-processing tasks such as sentence splitting, tokenization, part-of-speech tagging, and some form of parsing [41]. In the Natural Language Processing module, the OpenNLP is used to process the text through several phases in order to reach the desired form before entering the information extraction module as explained in the following sections.

3.2.1. Sentence boundary detection

The OpenNLP Sentence Detector is used to detect the end of a sentence by examining the punctuation character. The beginning of a sentence is assigned by the first non-whitespace character, and the last non-whitespace character is assumed to be then end of the sentence.

After detecting the sentence boundaries each sentence is written in a single line.

3.2.2. Tokenization

Finding parts of text is performed to implement additional processing on these tokens such as, stemming, stop words removal, lemmatization and converting text to lowercase [43,48]. The OpenNLP Tokenizers is used to segment each sentence to tokens based on whitespace characters [44].

3.2.3. Stop words removing

Stop words removing process is performed on the text generated from tokenization process to enhance the recognition of concepts and relations in the information extraction module A list of common stop words that frequently mentioned in physician's notes was created manually.

3.2.4. Part of speech tagging (POS)

POS Tagging is used in the NLP module to determine the tag of each word, then a set of rules is used to remove junk verbs from the sentences such as "started", "seeked". Removing these verbs will be useful in the information extraction module to concentrate on recognizing the noun phrases. The OpenNLP POS Tagger is used to define tokens with their corresponding word type using a probability model to predict the correct POS tag out of the tag set [44].

3.2.5. Sentence parsing

Noun phrasing is considered to be an important NLP technique used to investigate the possibility of combining traditional keywords in order to improve the quality of targeted information [17]. The parsing process starts by creating a parse tree for a textual unit that is a hierarchical data structure represents the syntactic structure of a sentence [48]. The OpenNLP parser is used to determine noun phrases in each sentence by extracting combination of words which tagged with NP (Noun Phrase). Medical concepts usually consisted of words (such as: fever, cough, toxemia, BMI, etc.) or phrases (such as: Nocturnal enuresis, loss of appetite, Acute sinusitis, etc.). In the information extraction module, the named entity recognition is applied to each noun phrase to choose the appropriate phrase that has a named entity type such as "shortness of breath". Fig. 4 shows a full example of the natural language processing module processes.

3.3. Information extraction module

The information extraction approach presented in this paper is based on a domain ontology. In our previous work, we developed a domain ontology called patient clinical data (PCD) [13]. PCD Ontology represents the clinical data produced during the healthcare activities through the patient visit. PCD ontology concepts is created by examining EHRs database, acquiring knowledge from physicians and domain experts, formal and informal text analysis of clinical terms, and integrating medical ontologies from the Unified Medical Language System UMLS (UMLS). The National Library of Medicine (NLM) provides several well-known 'knowledge infrastructure' resources such as UMLS Metathesaurus. UMLS Metathesaurus records synonyms and categories of biomedical concepts from numerous biomedical terminologies, which is useful in clinical NER [43]. The PCD ontology used to extract named entities from clinical notes. PCD ontology was formalized in the Ontology Web Language (OWL) using Protégé 5.2. We checked the logical consistencies and taxonomies classification in the PCD ontology automatically using HermiT 1.3.8.0.413 reasoner which reported valid ontology consistency and ontology taxonomy. According to ontology evaluation methods reported by [14], we used data-driven ontology evaluation and multiple-criteria evaluation approaches with the help of domain experts in medical informatics and clinical practice to evaluate the quality of the PCD ontology before using it in our proposed information extraction system. Following a data-driven evaluation approach, we investigated a collection of 250 clinical notes documents collected from three hospitals and five practices of different medical specialists. Sixty sample documents were chosen for manual analysis against PCD ontology. Also, we compared the EHRs data components and structure from three hospitals with the concepts, data properties and relationships of the PCD ontology. The evaluation results showed that PCD ontology has a good structure and coverage of clinical concepts by 87% of the investigated data, and all the required changes were done to improve the ontology concepts structure and coverage and to ensure the adequacy of the semantic categories of each concept. Finally, we used the multiple-criteria evaluation approach to asses PCD ontology against predefined criteria including correctness, completeness, clarity and conciseness. Manual evaluation on these criteria was conducted by fifteen domain experts (three internal medicine specialists, two pulmonology specialists, two clinicians from cardiology department, Professor of general surgery and oncology, three physicians from pediatrics department and four gynecology specialists). The domain experts confirmed that the classes, properties and individuals in PCD ontology are understandable, correctly represent essential aspects of the real healthcare workflow, fully describe the domain knowledge of patient clinical data in the EHR system and they complied that the ontology does not include irrelevant or redundant concepts. Fig. 5 shows a snapshot of the ontology. The high-level classes of the PCD ontology are: Patient, Encounter, Findings, Diagnosis, Procedures, Medications, Operations and Diagnostic tests. The patient concept means the person who has clinical problem and attend an encounter at the healthcare facility to solve this problem. Each patient has demographics and clinical data gathered through the activities of healthcare service.

The encounter class structures knowledge about the visit or admission or other contact between patient and health care provider. The findings class structures the information about the observation activities that physicians do during the patient encounter such as the history related to the patient complaints, the symptoms, the vital signs, the allergies, the immunizations, the general and the local examination. The diagnosis class defines the diseases or the conditions that the patients may have. It structures the information about the problem or the condition of the patient, the clinical statues, the severity of the disease as evaluated by the clinician, the start date of the disease and any additional notes related to the clinician decision. Procedures class consists of five subclasses which present the types of procedures performed by the patient as a part of treatment plan. Each procedure is determined by the clinician after the examination process and the diagnosis decision. Medications class structures the information about the A to Z pharmacological name of drug described by the clinical as a part of treatment plan, ingredients, interactions, and the instructions information about each drug. Operation class structures the information about each operation the patient performed, and the information related to the operation clinician stuff, operation equipments, preoperative and postoperative diagnosis, etc. Diagnostic tests class consists of radiology and laboratory tests. Diagnosis tests is performed as a part of observation process to assist in or confirm the diagnosis decision determined by the clinician.

3.3.1. Named entity recognition

Named Entity Recognition identifies specific words or phrases ('entities') and categorizes them as persons, locations, diseases, or medication. The common NER tasks are detection of entities and determining entities types [43,48]. There are a number of NER techniques available. Some use regular expressions and others are based on a predefined dictionary [48]. In this research, a predefined dictionary based on ontology concepts are used for recognizing named entities. Ontology concepts are organized in entities and relationships between entities. Each entity could be class, subclass, or instance of class. The PCD ontology represents the information model for OB-CIE system that model the clinical data produced through patient visit. PCD ontology

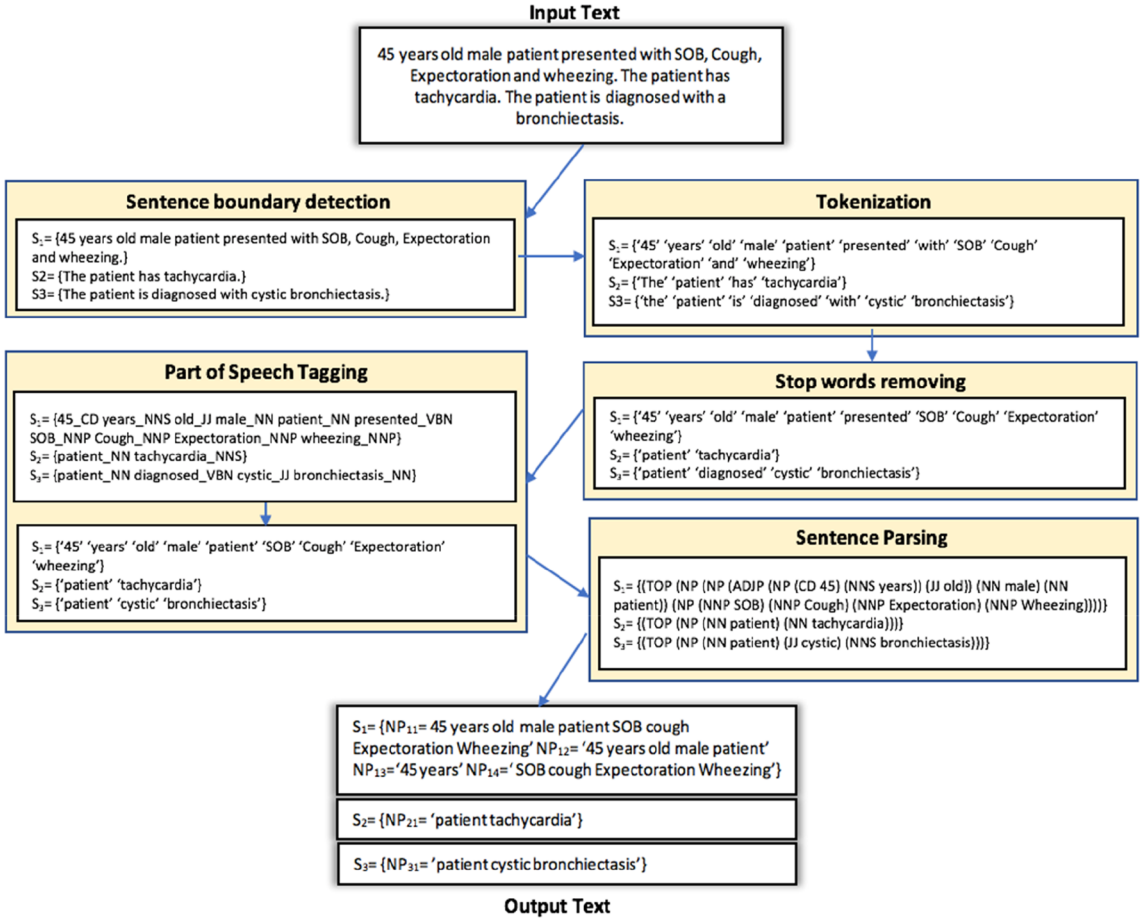

Fig. 4. The natural language processing module processes.

include three types of named entities, Entity (concept), Attribute (data property), Relation (object properties). Concept represents classes, subclasses, and instances (such as findings, patient medication), object properties represents the relationship between instances of two classes (such as Patient Has Disease), and data property represent an attribute of a specific class (such as "age" and "gender" are data property of Patient class). The suggested NER method in the OB-CIE system basically is done through two steps: construct a dictionary of semantic categories (named entities) and matching noun phrases to their semantic categories. LingPipe is used to implement the suggested NER method in the OB-CIE system. LingPipe provides an implementation of the Aho-Corasick algorithm [38]. The Aho-Corasick algorithm is processed to

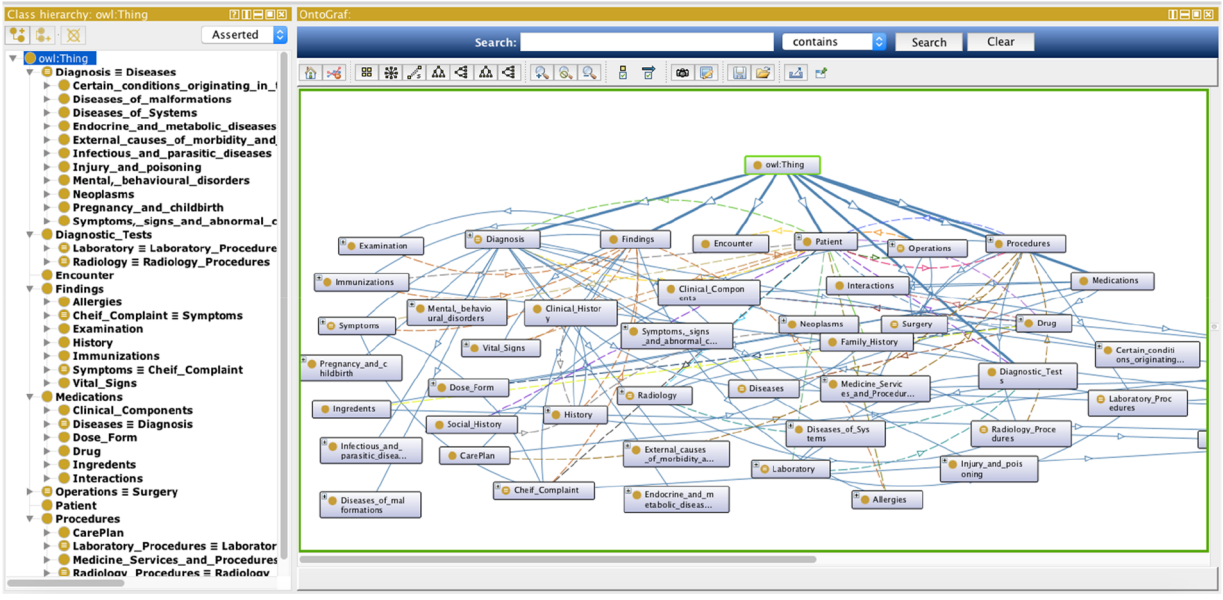

Fig. 5. A snapshot of the PCD ontology.

Table 1
Sample of the SPARQL query result.

Fact	Type	Class
Oedema	Owl:NamedIndividual	Symptoms
Laboratory_Procedures	Owl:Class	Procedures
Allergy_Onset	Owl:DataTypeProperty	Allergy
Asthma	Owl:NamedIndividual	Chronic_lower_respiratory_diseases
BMI	Owl:Class	Vital_Signs
Breast_Mammograohy	Owl:NamedIndividual	Radiology
Age	Owl:DataTypeProperty	Patient

find all matches of phrases against a dictionary in linear time independent of the number of matches or size of the dictionary [2].

3.3.1.1. Step1: Building a dictionary of semantic categories. Constructing a dictionary of the semantic categories is a procedure of mapping all concepts found in the PCD ontology to their parent classes. The dictionary consists of semantic groups and concepts categorized to their groups. This step begins with retrieving all facts in the ontology with their corresponding types and parent classes. To get all the facts presented in the PCD ontology, SPARQL was used. SPARQL is a query language used to formulate expressive queries over RDF data. Apache Jena is used to integrate the PCD ontology in the OB-CIE system and run SPARQL queries over the ontology. Apache Jena is an open source Java framework for building semantic web and Linked Data applications [32]. By using SPRQL, all the facts of the PCD ontology retrieved with their types (class, individual, object property, data property) and their parent classes. Table 1 shows sample of the query result. The retrieved data are organized to concepts and semantic groups with indication of their class hierarchy and type which can be entity or attribute. Table 2 shows some of the basic semantic categories extracted from PCD ontology which commonly mentioned in physician's clinical notes.

After organizing the data retrieved from the ontology to concepts and semantic categories, the dictionary of semantic categories is built from these entries. The dictionary entries comprise of two arguments: the semantic category (type) and the list of concepts belongs to it as shown in the following structure:

$$\begin{aligned}\text{Dictionary} = \{ &< Y_1, C_1, C_2,..,C_j,..C_m > \\ &< Y_2, C_1, C_2,..,C_j,..C_m > \\ &< Y_i, C_1, C_2,..,C_j,..C_m > \\ &..... \\ &< Y_n, C_1, C_2,..,C_j,..C_m > \}\end{aligned}$$

Y represents semantic groups; $Y = \{Y_1, Y_2,..,Y_i,..Y_n\}$ and C represents concepts $C = \{C_1, C_2,..,C_j,..C_m\}$. For recognizing values of attributes such as "45 years old" is a value of Patient "age" attribute and "twice daily" is a value of Drug "frequencyPerDay" attribute, a dictionary of predefined values is constructed based on investigating the most frequently mentioned values in physician's clinical notes through interviewing domain experts. Table 3 presents an example of some dictionary's entries of the frequently attributes values.

Table 2
The basic semantic categories extracted from PCD ontology.

Semantic category	Concepts
Patient[entity]	Patient
Diagnosis[entity]	pneumonia, toxemia, bronchiectasis
Examination ⇒ Findings[entity]	Tachycardia, NAD, rhonchi
Symptoms ⇒ Findings[entity]	Sneezing, Wheezing, Dyspnea
VitalSigns ⇒ Findings[entity]	Weight, temperature, BMI
Allergies ⇒ Findings[entity]	Penicillin, Mammals, Fowl Insulin
Immunizations ⇒ Findings[entity]	Influenza, DTP, Hepatitis B
SocialHistory ⇒ History ⇒ Findings[entity]	Smoker, heavy smoker
Drug ⇒ Medications[entity]	Crestor, Tritace, Concor, Plavix
DoseForm ⇒ Medications[entity]	Tablet, capsule, syrup
Laboratory ⇒ DiagnosticTests[entity]	CBC, INR, cytology
Radiology ⇒ DiagnosticTests[entity]	CT, endoscopy, US
Procedures[entity]	Breast biopsy, Hysterectomy

3.3.1.2. Step 2: Matching noun phrases to their semantic categories. After building the dictionary, the next step is to use a dictionary matcher to find all matches of phrases against the dictionary. The exacted dictionary chunker based on Aho-Corasick algorithmis is used for matching phrases to their semantic categories. The exact dictionary chunker extracts chunks based on exact matches of tokenized dictionary entries [9]. The process of matching concepts to their semantic categories starts after building the chunker. After that, All the chunk sets of the dictionary is returned with their character slice and the index of start and end character of character slice. Character slice of the sentences chunks are compared to the dictionary chunks to check the equality of them. By finding equal chunks, semantic group of each chunk is returned. The output of exact dictionary-based chunking process is the concepts and their semantic categories of the input sentences which written in patient clinical notes. By applying the named entity recognition process on the text outputted from the NLP module in Fig. 4, the three sentences will be chunking and comparing with chunk sets to return the semantic categories and the output will be concepts and semantic categories of these concepts. Fig. 6 presents the final output of the named entity recognition process.

As shown in the Algorithm 1, the sentences are partitioned into a set of chunks. Character slice is specified for each chunk with the index of start and end character of character slice.

Algorithm Concepts Classification

Inputs: Sentence (S) / Chunker TF
Output: Semantic Categories (Y) of Concepts (C)

```
Return the sentence chunking Sch
for each Sch ϵ S do
    Return Character Slice Cs
    Return start and end index of Cs
end
Return the chunk set of the chunking Cch
for each Cch ϵ chunk set do
    Return Character Slice Cs
    Return start and end index of Cs
end
for each Sch ϵ S do
    for each Cch ϵ chunk set do
        Compare Cs, start and index of Sch and Cch
        if Sch == Cch then
            Return semantic category Y of Sch
        end
    end
end
Return semantic categories Yof concepts C
```

Table 3
Example of some dictionary's entries of the frequently attributes values.

Semantic category	Example of concepts values
Age ⇒ Patient[Attribute]	12 years old, 45 years old
Gender ⇒ Patient[Attribute]	Male, female
MaritalStatus ⇒ Patient[Attribute]	Married, single
VitalSign_Value ⇒ VitalSigns ⇒ Findings[Attribute]	36.5C, 120/80, 50 kgs.
FrequencyPerDay ⇒ Drug ⇒ Medications[Attribute]	Twice daily, 3 time/day, 1x2, 1x3
TimeTaken ⇒ Drug ⇒ Medications[Attribute]	Before breakfast, after meals
Dose ⇒ Drug ⇒ Medications[Attribute]	600 mg, 500 mg, 100 ml
BodySite ⇒ Examination ⇒ Finidngs[Attribute]	Neck, chest, abdomen
Result ⇒ Examination ⇒ Finidngs[Attribute]	Fair, free
ConditionHistory_Onset ⇒ History ⇒ Findings[attribute]	2 weeks ago, few days ago

3.3.2. Sentences classification

A sentence or clinical note belongs to category or section inside the clinical notes document (e.g., Examination, History, Diagnosis). Rule-based classification is used to categorize sentences considering the semantic categories of concepts included in these sentences. By examining the main sections in physician's clinical notes, eleven sentence categories are determined to classify sentences. Table 4 presents sentence categories with category description and example of sentences belong to these categories. There are 22 semantic categories of concepts and 11 sentence categories, so, concepts categories will be mapped to sentence category in order to be used in sentence classification process. There are sentence categories equaled to one concept categories as Diagnosis equals to Diagnosis[entity] and there are sentence categories include many concept categories such as Medications equals to 5 concept categories, so they will be mapped to Medication category. The sentence classification rules are used to classify sentences according to classification of concepts. Firstly, the semantic categories of concepts belong to each sentence will be retrieved and mapping to sentence category. If all concepts have one category, the sentence will be classified to this category. If the concepts are classified to different categories, additional processing using rules will be performed to classify the sentence.

A weight of each sentence category will be specified based on the category of concepts. Then, the sentence category which have the maximum weight, will be the winner category and the sentence will be classified to it. For example, sentence "A female patient complains of nausea and vomiting" includes 4 concepts: (ICD10 [31]) "patient" will be ignored in sentence classification process to avoid affecting on classification result. (2) "female" is classified to (Patient) that increase weight of "patient" category by one (3) "nausea" and "vomiting" are classified to "Symptoms" that increase weight of "Symptoms" category by two. So, "Symptoms" category has the maximum weight and the sentence will be classified to it.

3.4. Recognizing extracted information in EHR database

The semantic categories were built based on the facts retrieved from the PCD ontology which represents the information model for the OBCIE system. The facts in the PCD ontology could be one of the following types:

- Entity-Class: represents the super class of concepts and attributes (such as medications or Drug class).
- Entity: represents the concepts that can be subclasses or individuals (such as Carvedilol).
- Relationship: represents the defined object properties between classes in the PCD ontology (such as Patient-Take-Drug).
- Attribute: represents the defined data properties of each class in the PCD ontology (such as Dosage).
- Attribute-Value: represents the possible data properties values for each class in the PCD ontology (such as 12.5 ml). A set of predefined values of attributes is constructed for class attributes based on investigating the most frequently mentioned values in physician's clinical notes through interviewing domain experts and analysis of clinical notes and patient records.

In the process of Named Entity Recognition, all the PCD ontology facts retrieved and organized to concepts and semantic groups with indication of their class hierarchy and type. Using a domain ontology in named entity recognition process enriches the value of the extracted information because it will not only extract the named entity type of the concept but also extract additional semantic information. For example, as shown in Table 3, concept "100 ml" belongs to semantic category of "Dose ⇒ Drug ⇒ Medications[Attribute]". This indicates the following information:

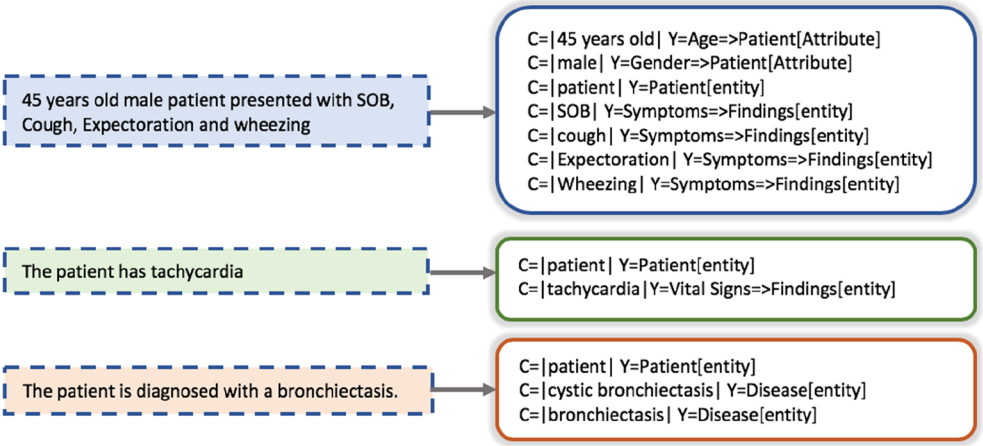

Fig. 6. The final output of the named entity recognition process.

Table 4
Description and examples of sentence categories.

Category	Description	Example
Patient	The patient information such as age and gender	A male patient is 30 years old
Symptoms	The patient's current complaint or reason for seeking medical care	The patient complains of nausea and vomiting
History	A detailed description of the growth of the patient's illness	The condition started 2 weeks ago by acute cough.
Social history	Marital status; dietary, sleep, and exercise patterns; use of coffee, tobacco, alcohol; diet, daily routine.	The patient is a cigar smoker Heavy smoker
Findings	The results of vital signs, general overview and systematic overview	The patient has pallor of skin/ Temperature is high
Examination	An evaluation of the body by means of sight, touch, and auscultation.	The patient has Palpitations/Swelling of breast
Disease (Diagnosis)	The nature and circumstances of a diseased condition	The patient has Eczema of LT nipple
Medications	A list of medications that contains drugs and drug instructions	Plavix 75 mg Tablet twice daily before meals
Procedures	Actions that may be used for disease prevention and patient safety	For Soft diet/ Recommendation for surgery
Laboratory	The diagnostic tests that include chemistry and hematologic labs.	CBC revealed leukocytosis neutrophilia
Radiology	The diagnostic tests that include radiology studies.	CT abdominal with and without contrast

- "[Attribute]" refer that "100 ml" is an attribute value.
- "Dose" is the attribute name of value "100 ml".
- "100 ml" is a Dose of Drug.
- "Dose" is an attribute of "Drug" class.
- "Medications" is the super class of "Drug" class.

The extracted information is represented as a set of "is a" relationships. Each relation represents an attribute or entity with its super class. To extract relationship between concepts to be recognized in the structured EHR database, we proposed a rule-based relation extraction method. The rule-based relation extraction method depends on the semantic categories of concepts and sentences extracted from the PCD ontology in the information extraction phase. The PCD ontology has a structure for clinical data represented in the EHR database, so, in the proposed relation extraction method, we used the rules for mapping between the PCD ontology structure and the EHR database structure. We investigated the clinical notes documents and derived that the concepts that fall in the same sentence are interrelated. For example, sentence" Plavix 75 mg twice daily before meals" classified to (Medications) category and includes 4 concepts classified as the following:

- Plavix:Drug ⇒ Medications[entity]
- 75 mg: Dose ⇒ Drug ⇒ Medications[Attribute]
- twice daily: FrequencyPerDay ⇒ Drug ⇒ Medications[Attribute]
- before meals: TimeTaken ⇒ Drug ⇒ Medications[Attribute]

The relation extraction rules were created based on the sentence category, concepts semantic categories and the structure of clinical data in the PCD ontology. To extract relationships, we used the sentence category to determine the higher category of the sentence concepts for indicating the class in the PCD ontology. For example, "Medication" class. The concepts semantic categories indicate that the sentence includes one entity and three attributes which all belongs to the same class "Medications" in the PCD ontology. So, the attributes are related to the entity. We can represent the relationships structured in the PCD ontology concepts as the following forms:

Form 1: Entity-name = Entity-value
Form 2: Entity-value. Attribute-name = Attribute-value

Form 1 is applied on the concepts whose semantic category ends with "[entity]" and form 2 is applied on the concepts whose semantic category ends with "Attribute". In the PCD ontology, the "Entity-name" represents an ontology class, "Entity-value" represents an individual, "Attribute-name" represents data property of an individual, "Attribute-value" represents a predefined value of data property of an individual. If we applied this relation forms in the previous example, the output will be:

Drug = "Plavix"
Plavix. Dose = "75 mg"
Plavix. FrequencyPerDay = "twice daily"
Plavix. TimeTaken = "before meals"

For the classified structures to be recognized by the structured database, a set of rules were generated manually for mapping PCD ontology structure to the EHR database structure. For example, entities in the Drug class is mapped to Drug source table in the EHR database, and data property of the Drug class is mapped to attributes in Drug table. Table 5 presents some attributes of EHR database with attribute source and example. When applying these mapping rules on the previous example, the output will be:

" Plavix 75 mg twice daily before meals" =≫ "Drug" source table.
"Plavix" =≫ "Drug_Code" attribute.
"75 mg" =≫ "Drug_Dose" attribute.
"twice daily" =≫ "Drug_FrequencyPerDay" attribute.
"before meals" =≫ "Drug_TimeTaken" attribute.

3.5. Evaluation methods

This paper focuses on free-text outpatient clinical notes that written by a physician from different medical specialists, such as cardiology and pediatrics. The OB-CIE system has been evaluated using 150 clinical documents represent 847 physician's clinical notes and 3165 concepts

Table 5
Some attributes of EHR database with attribute source and example.

Attribute	Source	Example
Patient_MRN	Patient	A 2234 patient
Disease_Code	Diagnosis	Pneumonia
Examination_Value	Examination	Basal crepitation
Symptom_Code	Symptoms	cough
VS_Code	VitalSigns	temperature
Allergy_Code	Allergies	Penicillin G
Immunization_Code	Immunizations	Fluvax
ConditionHistory_Code	History	Heavy smoker
Drug_Code	Drug	Zyrtec
Drug_DoseForm	Drug	Syrup
Laboratory_Code	Laboratory	CBC
Radiology_Code	Radiology	CT
Procedure_Code	Procedures	Soft diet
Age	Patient	5 years old
Gender	Patient	female
MaritalStatus	Patient	Married
VS_Value	VitalSigns	50 kgs.
Drug_FrequencypPerDay	Drug	Once daily
Drug_TimeTaken	Drug	Before breakfast
Drug_Dosage	Drug	200 mg
Examination_BodySite	Examination	Right chest
ConditionHistory_Onset	History	2 weeks ago

Table 6
Clinical documents description.

Clinical domain	Documents count (%)	Categories
Internal Medicine and Cardiology	30 (20%)	Complains, history of present illness, history of past illness, family history, general examination, local examination, Laboratory tests, Radiology tests, diagnosis, surgery, medications
General surgery and Oncology	30 (20%)	History, Examination, diagnosis, Laboratory tests, Radiology tests, recommendations, procedures, surgery, medications
Obstetrics and gynecology	20 (13.3%)	Vital signs, Diagnosis medications, laboratory tests, radiology tests, procedures, surgery
Pulmonology	40 (26.6%)	Symptoms, history, social history, examination, laboratory tests, radiology tests, diagnosis
Pediatrics	30 (20%)	Vital signs, diagnosis, medications

collected from 8 various healthcare facilities underlying 5 major medical specialists. The 150 clinical documents are chosen based on two criteria: the clarity of handwritten font and style, and the quality of document after scanning to the computer. Each clinical document describes an encounter with a physician and consists of handwritten free-text entries which belong to some of specified categories (Patient demographics, History, Symptoms, Vital signs, Examination, Laboratory tests, Radiology tests, Diagnosis, Procedures, Medications, etc.). Table 6 shows the number of documents and the clinical domain (medical specialists) of the collected clinical notes. The clinical notes documents were obtained from several sources such as clinics and hospitals, so, the documents structure is different and has several characteristics.

The document could be a white paper as shown in Fig. 7.a. It could have a sections or categories and the physicians write his/her notes inside each section as shown in Fig. 7b–d. In this case, the sections titles will be removed, and the handwritten text only will be processed. All physicians write clinical notes in English but some of them may write the drug dosage in Arabic numbers or letters. So, if the document has an Arabic text, it will be removed, and the English text only will be processed. Fig. 8 shows the OB-CIE information extraction from unstructured clinical notes written by a physician.

A list of medial domain experts consists of 5 physicians from 5 medical specialists (internal medicine and cardiology, general surgery and oncology, obstetrics and gynecology, pulmonology, and pediatrics) are considered as the gold standard, they analyzed the test set before testing in the OB-CIE system through four phases (pre-processing, NLP processing, Information extraction, and relation extraction): Firstly, the domain experts typed each concept inside the handwritten document in machine readable format to be used as the ground truth for comparison with the results outputted from the OCR process in the pre-processing module. The ground truth is used to determine the number of correctly recognized items (characters and words) and supply the total number of the ground truth items. Secondly, they specified noun phrases which consist of one or more clinical concept for comparison with the noun phrases outputted from the NLP module. The noun phrases extracted the NLP Sentence Parser were compared against the gold standard. The NLP Sentence Parser were evaluated using recall and precision measures. Recall was defined to be the number of noun phrases correctly identified, divided by the total number of actual noun phrases manually identified by a human domain expert while precision is the number of phrases correctly identified by the parser, divided by the total number of nouns identified by the parser. Thirdly, they classified noun phrases and sentences according to their specified categories. Finally, the domain experts extracted the relations from each document in the test set as a list of attributes to be stored in the patient record in the EHR system.

The testing results produced by OB-CIE system were compared against the results produced by the domain experts in each phase. A result that is extracted by both the domain experts and OB-CIE system is classified as true positive (TP), and a result that is extracted by OB-CIE system but not extracted by the domain experts is classified as false positive (FP). A result that is extracted by the domain experts but not extracted by OB-CIE system is categorized as false negative (FN).

To evaluate OB-CIE system, a standard evaluation metrics of Recall, Precision and F-measure were used. Precision and recall are defined in terms of true positive (TP), false positive (FP) and false negative (FN) as shown in the following equations:

$$Precision(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2PR}{P + R}$$

(TP + FP) refers to all items Extracted by the OB-CIE system, (TP + FN) refers to all the items found by the domain experts. Precision, recall and F-measure are calculated to evaluate the performance of the OB-CIE system modules.

4. Results and discussion

In this paper, we introduced an ontology-based clinical information system (OB-CIE) to extract clinical concepts from the unstructured clinical notes and convert them into structured data. The precision, recall and F-measure of the OB-CIE system have been evaluated using 150 clinical documents collected from various healthcare facilities. The evaluation results of the accuracy of OCR process in the OB-CIE system are shown in Table 7. The final ground truth data for evaluating OCR process consists of 22,881 characters and 4017 words. The overall evaluation of the OCR engine shows that the percentage of words in the original text correctly found by the OCR engine is 96.7% while the percentage of corrected characters is 98%, and the percentage of correctly found words with respect to the ground truth count of the OCR engine is 90.0% while the percentage of correctly found characters is 95.8%. All the misspellings in the recognized words are detected and corrected in the spell-checking process before moving to the NLP module. To evaluate the results of the NLP module in the OB-CIE system, we created a list of unique noun phrases identified after processing all the test documents in the NLP sentence parser to be compared against the gold standard. The domain experts identified 847 clinical notes or sentences and 1275 noun phrases containing 3165 concepts from the test set, while the NLP sentence parser extracted 1217 noun phrases. The Total number of noun phrases correctly identified by the NLP sentence parser was 1182. The evaluation results of the NLP sentence parser show that the overall recall value of noun phrases extraction is 95.4% and precision value is 97.1%.

The evaluation of concepts classification and sentence classification algorithms was conducted through two rounds: round one (R1) and round two (R2) on the same 150 clinical documents. The ground truth data for evaluating concepts and sentences extraction process consists of 847 classified sentences and 3165 classified concepts. The overall recall value of concepts classification in R1 is 83% and precision value is 88%. The evaluation results of R1 showed that there were some limitations in the concepts classification. Therefore, we examined the concepts that are incorrectly classified or not extracted to find the limitations and improve the system performance, then, we repeat the evaluation process on the same 150 documents which we called R2.

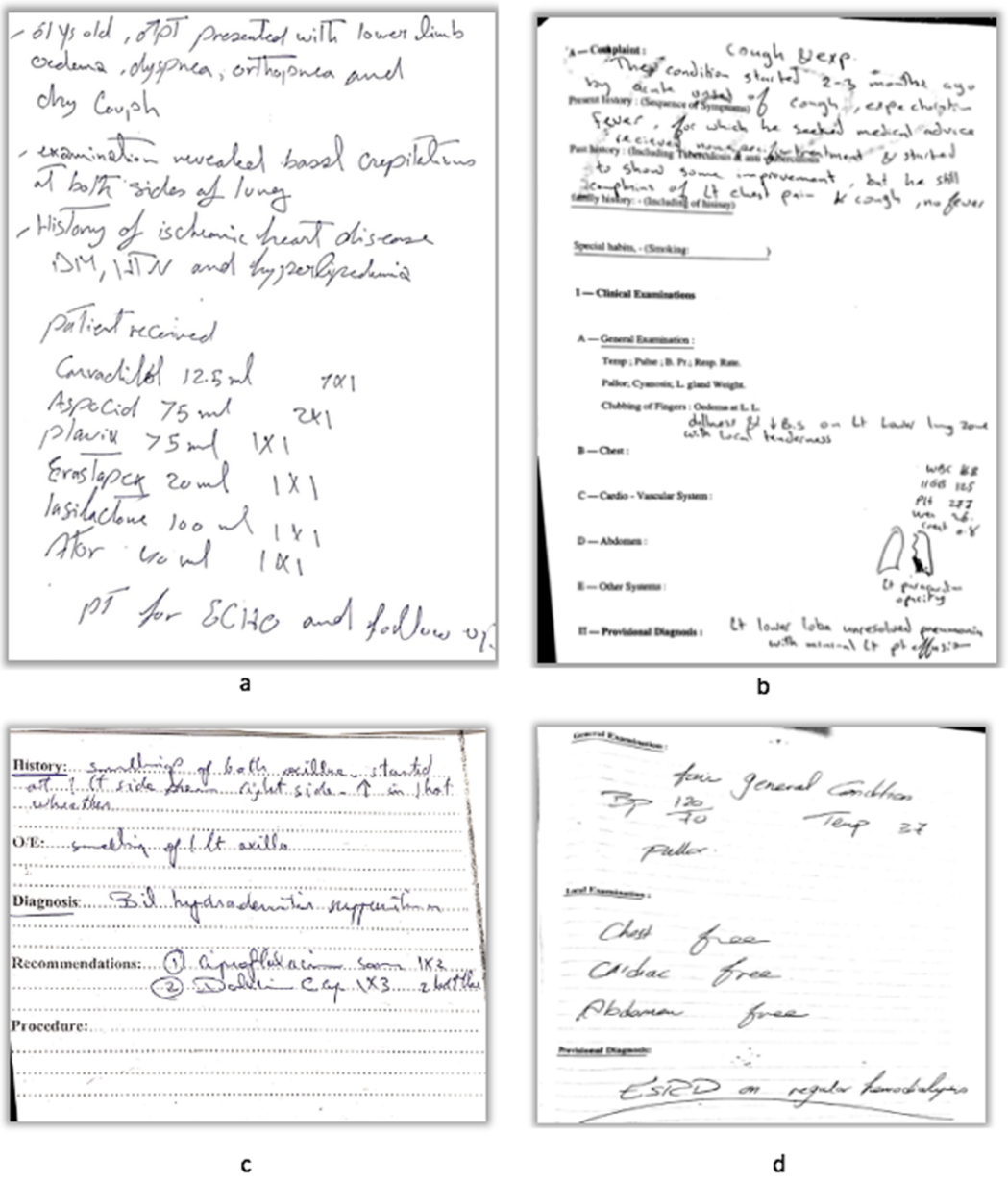

Fig. 7. Samples of clinical notes documents.

Table 6 gives the evaluation results of R1 and R2. In R2, the overall recall value of concepts classification is 96.7% and precision value is 99%. For overall system performance, the recall value is 94%, the precision value is 99%, and the F-measure is 96%.

By examining the concepts that are incorrectly classified or not extracted (unknown) in R1, we found some limitations. There are many concepts and abbreviations that are not identified in the PCD ontology. The PCD ontology is developed by integrating UMLS sources which are: SNOMED-CT, ICD10, LOINC, NDF-RT, RxNorm, and CPT. By searching UMLS data sources for the concepts that are not extracted, we noticed that they are identified in other UMLS sources. For example, “NAD” is an abbreviation of “no acute distress” identified in MEDCIN, “HTN” is an abbreviation of “Hypertensive disease” identified in CSP, MEDLINEPLUS, and NCI. We also found that there are many concepts that are not identified in the UMLS knowledge sources but are written by the physicians. For example, “Dipripam”, “Dalacin” and “Ator” are concepts written in physician’s clinical notes which should belong to “Drug ⇒ Medications” semantic category. To improve the performance of the OB-CIE system, some adjustments have been performed on the information extraction module. There are various studies performed for handling abbreviation recognition and de-ambiguous in clinical notes. Liu et al., presented a method for extracting abbreviations from the UMLS [39]. Xu et al., developed and evaluated several methods for detecting abbreviations from hospital admission notes [68]. Sheppard et al., performed a study to assess the frequency, nature and understanding of abbreviations in medical records [53]. Wu et al., presented a machine-learning methods for detecting Abbreviations in Discharge Summaries [66]. Moon et al., studied clinical acronyms and abbreviations using supervised machine-learning to understand issues related to practical clinical acronym and abbreviation [42]. Wu et al., described a hybrid system to normalize and encode clinical abbreviations [67]. Xu et al., proposed two neural word embedding features for the disambiguation of clinical abbreviations [69]. Kreuzthaler et al., presented an unsupervised approach for Abbreviation Detection in Clinical Narratives [36]. Wu et al., developed an open-source framework for clinical abbreviation recognition and disambiguation [65]. The UMLS Metathesaurus and Semantic Network are used to integrate concepts that are not identified (unknown) in PCD ontology. In the UMLS, each

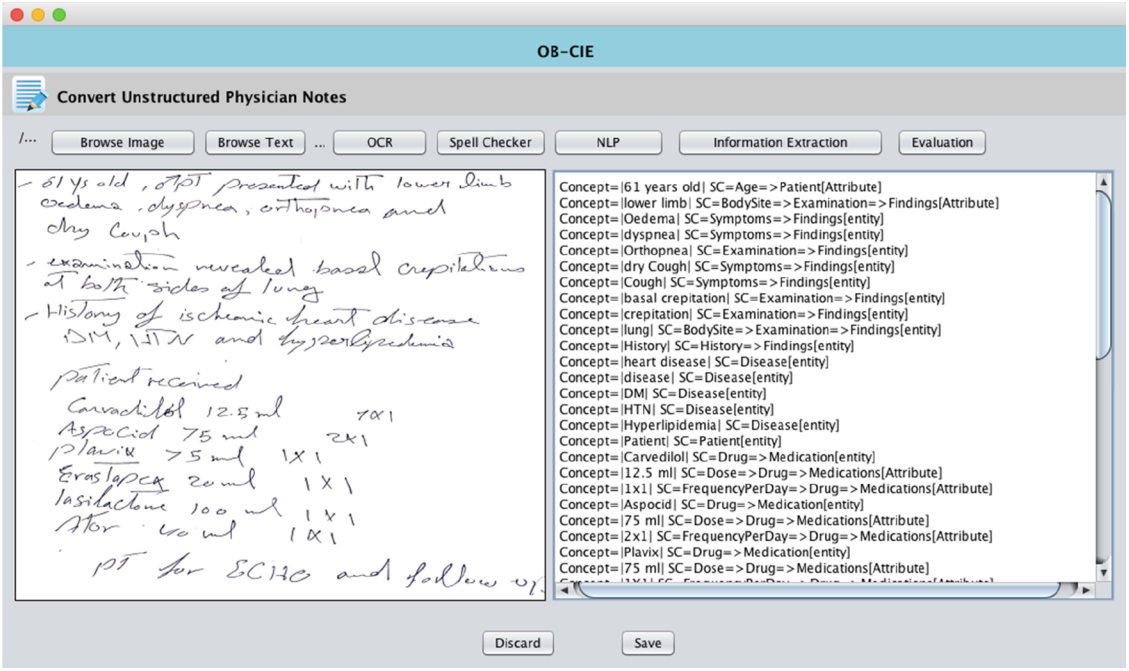

Fig. 8. The OB-CIE system.

Table 7
Evaluation results of the OCR engine.

Clinical domain	Character-Level					Word-Level				
	Ground Truth	TP	FP	Precision	Recall	Ground Truth	TP	FP	Precision	Recall
Internal Medicine and Cardiology	4869	4785	231	0.9539	0.9827	783	767	94	0.8908	0.9795
General Surgery and Oncology	5923	5862	275	0.9551	0.9897	972	945	110	0.8957	0.9722
Pulmonology	5247	5127	198	0.9628	0.9771	921	891	89	0.9091	0.9674
Obstetrics and Gynecology	3269	3197	114	0.9655	0.9779	659	620	39	0.9408	0.9408
Pediatrics	3573	3489	168	0.9540	0.9764	682	665	67	0.9084	0.9750
Total	22,881	22,460	986	0.9583	0.9808	4017	3888	399	0.9090	0.9670

Table 8
Evaluation results of R1 and R2.

Clinical domain	Concept-Level						Sentence-Level					
	Precision		Recall		F-measure		Precision		Recall		F-measure	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
Internal Medicine and Cardiology	0.899	1	0.811	0.921	0.852	0.958	0.958	1	0.921	0.966	0.939	0.982
General Surgery and Oncology	0.812	0.988	0.715	0.873	0.760	0.926	0.966	0.991	0.889	0.955	0.925	0.972
Pulmonology	0.874	0.958	0.813	0.891	0.842	0.923	0.895	0.963	0.891	0.958	0.892	0.960
Obstetrics and Gynecology	0.895	1	0.928	0.967	0.911	0.983	0.983	1	0.955	0.983	0.968	0.991
Pediatrics	0.931	1	0.889	0.915	0.909	0.955	1	1	0.973	0.973	0.986	0.986
Overall	0.882	0.989	0.831	0.913	0.855	0.949	0.960	0.991	0.925	0.967	0.942	0.978

Metathesaurus concept is categorized according to its semantic type. To integrate unknown concepts, the system will perform the following steps (ICD10) Search the UMLS Metathesaurus for the non-classified concept, (2) Get the concept CUI, Source, and Semantic type, (3) Match between UMLS semantic type and the PCD ontology classes, and (Tess4J) Insert the concept in its identified class. For matching between UMLS semantic type and the PCD ontology classes, each UMLS semantic type is converted to PCD class. For example, “Disease or Syndrome” is matched to “Diagnosis” class, “Sign or symptom” is matched to “Symptoms” class, etc. For concepts that are not identified in the UMLS Metathesaurus but written by the physicians, they are manually inserted into the PCD ontology classes according to the physician’s classification. For example, “Dipripam”, “Dalacin” and “Ator” are inserted to “Drug” class. Table 6 shows improvements in concept evaluation from R1 to R2. Finally, for overall system performance, the recall value is 94%, the precision value is 99%, and the F-measure is 96% (see Table 8).

Based on the output of the OB-CIE system and domain expert review of relations extraction, the gold standard from the 150 clinical notes documents contained 3165 attributes and 14 attribute sources. Table 9 shows the evaluation results of relation extraction process in terms of precision, recall and F-measure. For overall performance of the relation extraction process in the OB-CIE system, the recall value is 93%, the precision value is 95.5%, and the F-measure is 94.2%.

Table 9
Total numbers of attributes in each source, recall, precision and F-Measure values.

Attribute Source	Attributes	Precision	Recall	F-Measure
Medications	1177	0.9807	0.9932	0.9869
Examination	468	0.9890	0.9679	0.9784
Symptoms	376	0.9814	0.9840	0.9827
Diagnosis	298	0.9897	0.9765	0.9831
Laboratory	186	0.9347	0.9247	0.9297
Illness History	175	0.9649	0.9428	0.9537
Patient	159	0.9870	0.9559	0.9712
Procedures	87	0.9523	0.9195	0.9356
Vital Signs	86	0.9753	0.9186	0.9461
Radiology	74	0.8961	0.9324	0.9139
Social History	32	0.9354	0.9062	0.9206
Surgery	19	0.9444	0.8947	0.9189
Chef Complain	16	0.9333	0.875	0.9032
Family History	12	0.9090	0.8333	0.8695

Table 10
The performance of the OB-CIE system and Bushinak et al. in General surgery and Oncology domain.

Study	Method	Dataset	Named entities	Precision
Bushinak et al	Data mining	40	History, Examination, Diagnosis, Procedures	91.36%
OB-CIE	Ontology Hand-coded rules Exact-based dictionary	30	Patient, Examination, Diagnosis, Symptoms, Vital signs, Allergies, Immunizations, Social history, Drug, Dose form, Laboratory, Radiology, Procedures, Age, Gender, Marital status, Vital sign value, Drug frequency-per-day, Drug time-taken, Drug dose, Body site, Examination result, Condition History Onset	98.8%

The system performance in the general surgery and Oncology domain is compared with Bushinak et al. study [70]. Bushinak et al., developed a technique to convert the unstructured medical data to structured data without modifying the workflow of physicians using text mining and natural language processing techniques [70]. This study was chosen to be compared to our system because it has similar characteristics as the OB-CIE system. Bushinak et al. extracts the structured information from physician's handwritten clinical notes and converted to structured information for recognizing in EHR database as the OB-CIE system. The set of clinical notes the authors used in Bushinak et al method was collected from Prof. Hussien Bushinak clinic for general surgery and Oncology in Egypt, which is the same source of our 30 clinical notes documents in the general surgery and Oncology domain under supervision of Prof. Hussien Bushinak. The sample of these documents was shown in Fig. 7c. Table 10 shows the comparison between Bushinak et al and the OB-CIE system. Bushinak et al. achieved 91% precision value while the OB-CIE system achieved 98.8% precision in the same domain. The Authors used data mining to classify concepts while OB-CIE system used an ontology of patient clinical information. Bushinak et al. also, used just 4 named entities to classify concepts and any concept did not lie under these categories, was classified as unknown named entity. the OB-CIE systems achieved very high precision value when compared with Bushinak et al. system. The reason for this high score can be attributed to the usage of domain ontology in the information extraction instead of data mining which indicate that ontology is a very important component of our system. The PCD ontology is the basic information model for our system which includes concepts and structure of all the patient data produced through the healthcare activities, and thus, can further be used to cover all the required named entities for classifying concepts in patient clinical notes. Using the PCD ontology in our information extraction system makes the patient data more valuable and can be used to assist in the decision-making process. In our future work, we will use the semantic web rules for reasoning and inferring new knowledge from the PCD ontology which can help in medical researches and improve the healthcare process.

Contributors

All authors listed on the manuscript have contributed sufficiently to the project to be included as authors.

Funding

The author(s) received no specific funding for this work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Ajami, T. Bagheri-Tadi, Barriers for adopting electronic health records (EHRs) by physicians, Acta Inform. Med. 21 (2) (2013) 129.
[2] V. Alfred, Complexity, Algorithms for finding patterns in strings, Algorithms, Complexity 1 (2014) 255.
[3] G. Antoniou, E. Franconi, F. Van Harmelen, Introduction to semantic web ontology languages, Reasoning web, Springer, 2005, pp. 1–21.
[4] D.E. Appelt, Introduction to information extraction, Ai Commun. 12 (3) (1999) 161–172.
[5] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: Paper presented at the Proceedings of the AMIA Symposium, 2001.
[6] A.R. Aronson, F.M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (3) (2010) 229–236.
[7] A.M. Association, Improving Care: Priorities to Improve Electronic Health Record Usability, American Medical Association, Chicago, IL, 2014.
[8] Association, H. F. M., Overcoming barriers to electronic health record adoption. Results of survey and roundtable discussions conducted by the Healthcare Financial Management Association, 2006.
[9] B. Baldwin, K. Dayanidhi, Natural language processing with Java and LingPipe Cookbook, Packt Publishing Ltd., 2014.
[10] R.J. Baron, E.L. Fabens, M. Schiffman, E. Wolf, Electronic health records: just around the corner? Or over the cliff? Ann. Int. Med. 143 (3) (2005) 222–226.
[11] A. Boonstra, M. Broekhuis, Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions, BMC Health Services Res. 10 (1) (2010) 231.
[12] A. Boonstra, A. Versluis, J.F. Vos, Implementing electronic health records in hospitals: a systematic literature review, BMC Health Services Res. 14 (1) (2014) 370.
[13] H. Boshnak, S. AbdelGaber, AmanyAbdo, E. Yehia, Ontology-Based Knowledge Modelling for Clinical Data Representation in Electronic Health Records, Int. J. Comput. Sci. Inform. Sec. 16 (10) (2018) 68–86.
[14] J. Brank, M. Grobelnik, D. Mladenic, A survey of ontology evaluation techniques, in: Paper presented at the Proceedings of the conference on data mining and data warehouses (SiKDD 2005), 2005.
[15] P. Carayon, T.B. Wetterneck, B. Alyousef, R.L. Brown, R.S. Cartmill, K. McGuire, J.M. Walker, Impact of electronic health record technology on the work and workflow of physicians in the intensive care unit, Int. J. Med. Inf. 84 (8) (2015) 578–594.
[16] M. Cheriet, N. Kharma, C.-L. Liu, C. Suen, Character Recognition Systems: A Guide for Students and Practitioners, John Wiley & Sons, 2007.
[17] G. Chowdhury, Natural language processing, Ann. Rev. Inform. Sci. Technol. 37 (1) (2003) 51–89.
[18] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, P.C. De Groen, Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model, J. Biomed. Inform. 42 (5) (2009) 937–949.
[19] C. Cotea, Electronic health record adoption: perceived barriers and facilitators, Research Coordination Unit, CMVH, 2010.
[20] L. Cui, A. Bozorgi, S.D. Lhatoo, G.-Q. Zhang, S.S. Sahoo, EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification, Paper presented at the AMIA Annual Symposium Proceedings, (2012).
[21] S. Doan, M. Conway, T.M. Phuong, L. Ohno-Machado, Natural language processing in biomedicine: a unified system architecture overview, Clinical Bioinformatics,

Springer, 2014, pp. 275–294.
[22] G. Fette, M. Ertl, A. Wörner, P. Kluegl, S. Störk, F. Puppe, Information Extraction from Unstructured Electronic Health Records and Integration into a Data Warehouse, in: Paper presented at the GI-Jahrestagung, 2012.
[23] R.M. Frankel, EHR and Physician–Patient Communication, in: Safety of Health IT, Springer, 2016, pp. 129–141.
[24] M. Friedberg, F. Crosson, M. Tutty, Physicians' concerns about electronic health records: implications and steps towards solutions, Health Affairs Blog 11 (2014) 963.
[25] C. Friedman, P.O. Alderson, J.H. Austin, J.J. Cimino, S.B. Johnson, A general natural-language text processor for clinical radiology, J. Am. Med. Inform. Assoc. 1 (2) (1994) 161–174.
[26] M.-P. Gagnon, E.K. Ghandour, P.K. Talla, D. Simonyan, G. Godin, M. Labrecque, M. Ouimet, M. Rousseau, Electronic health record acceptance by physicians: testing an integrated theoretical model, J. Biomed. Inform. 48 (2014) 17–27.
[27] M.-P. Gagnon, D. Simonyan, E.K. Ghandour, G. Godin, M. Labrecque, M. Ouimet, M. Rousseau, Factors influencing electronic health record adoption by physicians: A multilevel analysis, Int. J. Inf. Manage. 36 (3) (2016) 258–270.
[28] V. Garla, V.L. Re III, Z. Dorey-Stein, F. Kidwai, M. Scotch, J. Womack, A. Justice, C. Brandt, The Yale cTAKES extensions for document classification: architecture and application, J. Am. Med. Inform. Assoc. 18 (5) (2011) 614–620.
[29] H. Harkema, I. Roberts, R. Gaizauskas, M. Hepple, Information extraction from clinical records, in: Paper presented at the Proceedings of the 4th UK e-Science All Hands Meeting, 2005.
[30] D. Heisey-Grove, V. Patel, Physician Motivations for Adoption of Electronic Health Records, Office of the National Coordinator for Health Information Technology, Washington (DC), 2014.
[31] ICD10, International Classification of Diseases, Version 10 (ICD10), 2017. Retrieved from https://bioportal.bioontology.org/ontologies/ICD10.
[32] A. Jena, A free and open source Java framework for building Semantic Web and Linked Data applications, 2018. Retrieved from https://jena.apache.org.
[33] S.B. Johnson, S. Bakken, D. Dine, S. Hyun, E. Mendonça, F. Morrison, T. Bright, T. Van Vleck, J. Wrenn, P. Stetson, An electronic health record based on structured narrative, J. Am. Med. Inform. Assoc. 15 (1) (2008) 54–64.
[34] JOrtho, Java spell-checking library, 2018. Retrieved from http://jortho.sourceforge.net.
[35] J. Kozák, M. Necaský, J. Pokorný, Extracting Medical Information Using Linked Data, in: Paper presented at the SWAT4LS, 2012.
[36] M. Kreuzthaler, M. Oleynik, A. Avian, S. Schulz, Unsupervised abbreviation detection in clinical narratives, in: Paper presented at the Proceedings of the clinical natural language processing workshop (ClinicalNLP), 2016.
[37] R. Leaman, R. Islamaj Doğan, Z. Lu, DNorm: disease name normalization with pairwise learning to rank, Bioinformatics 29 (22) (2013) 2909–2917.
[38] LingPipe, Tool kit for processing text using computational linguistics, 2018. Retrieved from http://alias-i.com.
[39] H. Liu, Y.A. Lussier, C. Friedman, A study of abbreviations in, *the UMLS.* Paper presented at the Proceedings of the AMIA Symposium, (2001).
[40] M. Lyman, N. Sager, E.C. Chi, L.J. Tick, N.T. Nhan, Y. Su, F. Borst, J. Scherrer, Medical Language Processing for Knowledge Representation and Retrievals, in: Paper presented at the Proceedings. Symposium on Computer Applications in Medical Care, 1989.
[41] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearbook Med. Inform. 17 (01) (2008) 128–144.
[42] S. Moon, S. Pakhomov, G.B. Melton, Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations, Paper presented at the AMIA annual symposium proceedings, (2012).
[43] P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman, Natural language processing: an introduction, J. Am. Med. Inform. Assoc. 18 (5) (2011) 544–551.
[44] OpenNLP, Apache OpenNLP, 2018. Retrieved from https://opennlp.apache.org.
[45] J. Piskorski, R. Yangarber, Information extraction: Past, present and future, Multi-source, Multilingual Information Extraction and Summarization, Springer, 2013, pp. 23–49.
[46] L. Poissant, J. Pereira, R. Tamblyn, Y. Kawasumi, The impact of electronic health records on time efficiency of physicians and nurses: a systematic review, J. Am. Med. Inform. Assoc. 12 (5) (2005) 505–516.
[47] S.E. Pollard, P.M. Neri, A.R. Wilcox, L.A. Volk, D.H. Williams, G.D. Schiff, H.Z. Ramelson, D.W. Bates, How physicians document outpatient visit notes in an electronic health record, Int. J. Med. Inf. 82 (1) (2013) 39–46.
[48] R.M. Reese, Natural Language Processing with Java, Packt Publishing Ltd., 2015.
[49] RXNORM, Normalized names for clinical drugs, 2017. Retrieved from https://bioportal.bioontology.org/ontologies/RXNORM.
[50] G.K. Savova, J. Fan, Z. Ye, S.P. Murphy, J. Zheng, C.G. Chute, I.J. Kullo, Discovering peripheral arterial disease cases from radiology notes using natural language processing, Paper presented at the AMIA Annual Symposium Proceedings, (2010).
[51] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.
[52] C. Seebode, M. Trautwein, M. Ort, J.-M. Lehmann, A clinical information management platform for semantic exploitation of clinical data, in: Paper presented at the Proc. International MultiConference of Engineers and Computer Scientists, 2013.
[53] J.E. Sheppard, L.C. Weidner, S. Zakai, S. Fountain-Polley, J. Williams, Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping, Arch. Disease Childhood 93 (3) (2008) 204–206.
[54] J. Shu, Free text phrase encoding and information extraction from medical notes, Massachusetts Institute of Technology, 2005.
[55] B. Smith, W. Ceusters, An ontology-based methodology for the migration of biomedical terminologies to electronic health records, Paper presented at the AMIA Annual Symposium Proceedings, (2005).
[56] R. Smith, An overview of the Tesseract OCR engine, in: Paper presented at the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007.
[57] E. Soysal, I. Cicekli, N. Baykal, Design and evaluation of an ontology based information extraction system for radiological reports, Comput. Biol. Med. 40 (11) (2010) 900–911.
[58] E. Soysal, I. Cicekli, N. Baykal, Design and evaluation of an ontology based information extraction system for radiological reports, Comput. Biol. Med. 40 (11–12) (2010) 900–911.
[59] Tess4J, JNA wrapper for Tesseract OCR, 2018. Retrieved from http://tess4j.sourceforge.net.
[60] M. Toepfer, H. Corovic, G. Fette, P. Klügl, S. Störk, F. Puppe, Fine-grained information extraction from German transthoracic echocardiography reports, BMC Med. Inf. Decis. Making 15 (1) (2015) 91.
[61] H. Townsend, Natural language processing and clinical outcomes: the promise and progress of NLP for improved care, J. AHIMA 84 (2) (2013) 44–45.
[62] UMLS, The Unified Medical Language System (UMLS), 2017. Retrieved from https://uts.nlm.nih.gov/metathesaurus.html.
[63] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: a literature review, J. Biomed. Inform. 77 (2018) 34–49.
[64] A. Wright, E.S. Chen, F.L. Maloney, An automated technique for identifying associations between medications, laboratory results and problems, J. Biomed. Inform. 43 (6) (2010) 891–901.
[65] Y. Wu, J.C. Denny, S. Trent Rosenbloom, R.A. Miller, D.A. Giuse, L. Wang, C. Blanquicett, E. Soysal, J. Xu, H. Xu, A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD), J. Am. Med. Inform. Assoc. 24 (e1) (2016) e79–e86.
[66] Y. Wu, S.T. Rosenbloom, J.C. Denny, R.A. Miller, S. Mani, D.A. Giuse, H. Xu, Detecting abbreviations in discharge summaries using machine learning methods, Paper presented at the AMIA Annual Symposium Proceedings, (2011).
[67] Y. Wu, B. Tang, M. Jiang, S. Moon, J.C. Denny, H. Xu, Clinical Acronym/Abbreviation Normalization using a Hybrid Approach, Paper presented at the CLEF (Working Notes), (2013).
[68] H. Xu, P.D. Stetson, C. Friedman, A study of abbreviations in clinical notes, Paper Presented at the AMIA Annual Symposium Proceedings, (2007).
[69] J. Xu, Y. Zhang, H. Xu, Clinical abbreviation disambiguation using neural word embeddings, Paper presented at the Proceedings of BioNLP, 2015, p. 15.
[70] H. Bushinak, S. AbdelGaber, F.K. AlSharif, Recognizing the electronic medical record data from unstructured medical data using visual text mining techniques, Int. J. Comput. Sci. Inf. Secur. 9 (6) (2011) 25–35.