The Use of Ontology in Clinical Information Extraction

S Jusoh¹, A Awajan¹ and N Obeid³
¹School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan
²School of Information Technology, Jordan University, Amman, Jordan
³s.ibrahim@psut.edu.jo

Abstract. Extracting clinical data from medical or clinical reports is a crucial effort. These records contain the most valuable pieces of evidence of treatments in humans. Integration of information extraction (IE) and ontology can produce a great tool for clinical concept extraction. The aim of this paper is to present a quick overview of the research work which has applied IE and ontology approaches in medical or clinical concepts extraction. This paper also presents our proposed framework based on the integration of both approaches mentioned above for extracting clinical concepts.

1. Introduction
Medical or clinical records are one of the most valuable sources of information and substantiation of treatments in individuals. These records contain important items such as eligibility criteria, a summary of diagnose results, and prescribed drug which are normally recorded in unstructured free text. Extracting medical or clinical information from health records is a critical task in these days, especially with the adoption of the electronic health records. These records are normally stored in texts form, contain valuable unstructured data that may give insights for a better decision-making process. Gaining insight from a tremendous amount of unstructured clinical data has been a critical and challenging issue for medical organizations such as hospitals. Having an automated system that is able to read patients’ medical reports, extract medical entities, and analyze the extracted data, using the stored knowledge, and presenting the analyses to clinicians/users in a visualized form is not only very desirable but also a necessity. The challenging part is how to extract and encode the unstructured data to improve an overall healthcare system. Information extraction (IE) is a potentially good answer. There are three major elements that should be extracted from these clinical records: entities, attributes, and relations between them [1].

The first stage in exploring the unstructured data involves utilizing information extraction (IE) methods and techniques on texts. IE is one of natural language processing (NLP) applications that focuses on extracting valuable information from unstructured data. The aim of IE is to turn unstructured data which are in the form of free texts into structured representations [2]. IE a source of structured representations creation for a selected piece of information from free texts. It is a process of finding particular data/information in human language texts by dealing with the extraction of particular name entities, relationships and events, in the texts. Thus, the IE task is established by its input and its output (a target of extraction) and is also considered as a pre-processing technique of a text [3].
Technically, IE is an NLP system that recognizes and extracts information such as objects, events, and relationships among the objects from texts [4]. There are two fundamental tasks in IE, (1) named entity recognition and (2) relation extraction. A named entity is a token of words which recognizes a real-world entity such as "Jordan", "United States" and so on. Named entity recognition is an operation of locating and classifying entities according to their classes such as a person, an organization, a location and so on. However, while being considered as unstructured data, text data contain various internal structures that are pertinent to linguistics, pragmatic, and information. There are some existing data analysis tools such as statistical analysis tools and/or structured query language tools. However, these tools are unable to perform well due to the complexity of a human language. It is also worth to mention here that processing natural language is not a single monolithic procedure. For instance, in entity extraction, string matching does not always work for 2 main reasons: (1) dictionaries are always incomplete and (2) named entities are usually dependent on the context of use [5]. The second fundamental task in IE is relation extraction. It is a process of finding and locating semantic relations among entities. There are variety of techniques that have been used for relation extraction [5]. The most commonly used are classification techniques for finding relations between entities which co-occur in a sentence. However, the major challenge here is a possibility for having a relation between two entities span across multiple sentences.

Ontology, on the other hand, is a knowledge representation mechanism that specifies a formal definition for a set of commons terms in a particular domain [6] or common-sense knowledge [7] about the real world. These terms are normally known as classes or concepts and the relationships between the concepts are called as

Since IE is concerned with extracting entities and relations for a specific domain and an ontology is a way of representing concepts and relationships, combining both of these enhance the process of extracting unstructured data from texts. This has led to the emergence of ontology-based information extraction (OBIE) as a subfield of IE [8]. In OBIE systems, data extraction is guided by ontologies and the extracted data could be meaningfully represented using ontologies [8].

The aim of this paper is to touch on some aspects of medical information extraction research works and to propose a theoretical framework on ontology-based clinical information extraction. The paper is organized as follows. Section 2, presents some of the related work. Section 3 presents our proposed clinical based-information extraction (COBIE) framework. Section 4 presents the conclusion.

2. Some related work
Text is the predominant explicit knowledge for information sharing and exchange among experts. In the medical field, medical records are normally used among experts in the field to store, exchange or share critical information about patients. However, the textual form of these notes and their volume makes it difficult to survey even a single patient’s complete record. In a clinical task, expressing clinical events and entities are more natural and expressive if it is in the form of free text. Extracting and encoding clinical entities and events has gained great attention from NLP researchers and is still an ongoing active research area [9].

Clinical or medical entity extraction is a basis for higher-level tasks in health information retrieval such as illness diagnosis, illness prediction and so on. Clinical named entity recognition is normally involved with three types of classes: problems, tests, and treatments [9]. For instance, subclasses of class Problem constitutes names of concepts such as syndrome or disease, anatomic or abnormality, sign or symptom and so on. Subclasses of class Test constitute names of concepts such as a laboratory or diagnostic procedure, while subclasses of class Treatment constitute concepts such as drug, procedure, medical device, and so on.

A common approach to IE is to use patterns that define lexical, syntactic and semantic constraints. The patterns and rules are applied to texts through various levels of linguistic analysis, such as phrase chunking and full syntactic parsing [10]. Authors in [11] have used a rule-based abbreviation for mapping to their full forms in medical text.
Authors in [12] proposed keyphrases method to extract medical concepts such as DNA, and its association such as "nucleic acid", "biological information", and "proteins", from a set of 3000 medical abstracts. The method consists of 3 steps: document preprocessing, noun expansion and ranking candidate keyphrases. Authors in [13] proposed Intelligent Clinical Notes System, a system which was able to retrieve patient notes and extract useful information from the notes. This system focuses on only the electronic records. The use of rules on clinical entity extraction requires knowledge of human experts to define rules and expressions. In the case, where the expert knowledge is not available, machine learning approaches are the alternatives. Machine learning approaches which have been used for medical concepts extraction include Support Vector Machine (SVM), Conditional Random Fields (CRF) [14], Hidden Markov Models (HMM) [15] and Maximum Entropy Models (MEM) [16]. Machine learning techniques are normally trained on the available annotated clinical data sets. According to [9], the most recent tools of medical named entity recognition are based on machine learning approaches such as conditional random fields. The work in [17] identified the name of symptoms, diseases and drugs from Chinese electronic medical records by utilizing Cascaded Conditional Random Field approach.

The author in [1] presented a distant supervision approach to generate training data set. In this approach, texts are aligned to a knowledge base, and learned to extract relations between a medical entity and its corresponding attribute(s). The syntactical structure of a sentence was used to indicate the mapping direction. The authors in [9] stated that the aim of medical information extraction is to locate phrases which refer to medical concepts where they have experimented with three concept extraction models including MetaMap [18], machine learning approach using sequential classifier(s) Conditional Random Fields (CRF) [14]. It was reported that the combination of CRF and MetaMap features produces better results.

Authors in [19] performed the medical feature extraction method for Psoriasis disease using a machine learning approach. The authors in [20] performed IE from human handwritten documents in diabetes chronic disease based on 7 keywords parameters. Their IE system follows the following procedure, tokenization, filtering, stemming, weighting, and finding similarity. SVM model was performed before conducting a similarity procedure.

Ontologies are classified in many ways using abstraction and application criteria. Based on the criteria, ontologies exist in many forms. Authors in [21] classified ontologies into 3 main levels as shown in Figure 1 below. An application ontology defines concepts according to a particular domain application. For instance, an ontology for types of diseases falls into the application ontology.

![Figure 1. Level of ontologies [21]](image)

In text mining applications, ontologies are categorized into domain ontologies and common ontologies. A common ontology normally begins with an epistemic view and defines the relationships among general objects. A common ontology approach is normally based on the typical semantic dictionary such as English WordNet [22]. Researchers have put efforts in text mining study by
developing text mining applications based on a domain ontology. For example, reference [23] used text mining to learn the target ontology from text documents and then used the same target ontology to categorize texts. They reported by using the ontology, both supervised and unsupervised text categorization approaches are improved.

In a domain ontology, a vocabulary is used to represent classes and conceptions which constitutes the basic elements of knowledge about the domain. The relationships of a domain ontology depend on words to connect, so the vocabulary is the key to construction in a domain ontology [33]. To prepare the generic system for an actual application, a domain-specific ontology must be modeled, to store the structural information into the metadata model. Therefore, the user defines what data entities exist in the identified domain (Patient, Disease_name and Treatment_type) what kind of attributes they have such as Disease.DiagnoseDate and Patient.DateOfBirth, and in what kind of relationships they are such as a patient can sick with various diseases and receive different types of treatments [34].

Ontology has been considered as one of the visualization techniques which can be used to present extracted medical information from unstructured texts (medical reports) in a structured manner. The salient feature of ontology is that it can give patterns of semantics from free texts. Authors in [24] presented a model of medical concepts extraction using a linguistic ontology as presented in Figure 2. The aim of the model is to address the text mining applications for the domain. Despite the fact that the number of research work on medical ontologies is limited, some results have been reported in the literature, and it is worth to mention that most of the published research work normally focuses on constructing automated ontologies rather than utilizing them for the purpose to be used by clinicians.

According to [25], ontology can be constructed according to the type of knowledge resource such as structured texts, unstructured texts, plain texts, and so on. They proposed an automatic ontology construction based on the existing ontology for gastrointestinal diagnosis. Medical ontologies are normally developed for specified problems within the medical domain. For example, reference [26] reported on using an ontology to represent medical data entities for an intensive surgical care problem. The reported ontology describes a set of activities in a surgical intensive care such as cardiovascular, traumatology, general surgery, obstetrics, neuro-traumatology, and neurosurgery. Authors in [27] used ontologies to arrange and describe medical concepts. Reference [28] used Bayesian networks (BNs) to extract DFG triplets which consist of diseases, food, and genes from unstructured text data and a large number of PubMed abstracts. In this work, the ontology concepts were used inside the Bayesian networks, where each concept in the ontology is represented as a node in the Bayesian network.

Reference [29] reported research work on creating pneumology ontology from patients’ discharge summaries. Authors in [30] used an ontology to identify and extract targeted information from two medical books on prescriptions and febrile disease, respectively. An ontology for the Nuclear Cardiology has been reported in [31]. Their work was conducted based on Unified Medical Language System (UMLS). UMLS is a set of files and software which connect health and biomedical vocabularies and standards to facilitate interoperability among systems.

An ontology that focuses on malaria diseases had been reported in [32]. The approach adopted for creating the ontology was quantitative and manually-driven given that the data needed for the
development of the ontology was written as it was being developed. The Protégé OWL software was used to create malaria ontology.

In OBIE, IE process utilizes an ontology to extract desired concepts and present them as an output in an ontology form. As previously mentioned, IE is actually concerned with the process of extracting information from a specific domain, a formal and specific specification of concepts within the domain that are presented in an ontology form can be very practical to this process. For instance, a geopolitical ontology that defines concepts such as cities, states and countries can be utilized to guide the IE process. This is the main idea behind OBIE.

3. Clinical ontology-based information extraction (COBIE)

We have noted that there is no evidence reported in the literature for creating an ontology for pediatric diseases. We aim in this paper to develop a system which will be used to help clinicians in making appropriate decisions. Thus, there is an urgent need to conduct a research study that is able to store tacit and explicit medical knowledge related to pediatric diseases to support clinicians in diagnosing patients and making appropriate decisions. This type of systems will be able to support a limited number of clinicians to discover these diseases from a patient’s medical records at an early stage.

The proposed system should be able to help in extracting and giving a visual representation of medical data extracted from written medical texts such as a patient’s medical history reports. The ontology approach is seen as the most appropriate method for knowledge representation and visualization. As stated in [25], medical ontology construction in the real world remains challenging.

In this research, we view an ontology as a directed graph consists of nodes and edges. Nodes in the directed graph represent medical or clinical concepts while edges describe relationships between them. Two nodes which are connected through an edge are known as a triple. The procedures can be expressed through three kinds of ontologies (1) a source ontology, (2) a target ontology, and (3) a mapping ontology. The source ontology corresponds to the extraction of clinical concepts, whereas a target ontology corresponds to the loading process. The source ontology behaves as an inventory of all available clinical name entities in the source system. A mapping ontology is used to connect between a source ontology and a target ontology. Later, the mapping ontology can be automatically translated into a visualization technique. The adoption of mapping and merging ontologies to perform an automatic ontology was also presented in [35].

There are 3 major phases in conducting this study, creating a source ontology, constructing a target ontology, and producing a mapping ontology as shown in Figure 3.

![Figure 3. Proposed framework of clinical ontology-based IE (COBIE).](image)

The first phase includes extracting concepts or entities and relationships between them which are relevant to the paediatric diseases from a patient’s medical report. In this phase, information extraction
techniques such as tokenization and named entity tagging will be deployed. During this process, we may need to customize the existing IE techniques to efficiently and accurately extract relevant entities, concepts, and relationships. The source ontology is stored as a graph database. In creating a source ontology, it is necessary for a specialist paediatric (a human expert) to intervene by evaluating and verifying the accuracy of knowledge for a source ontology. The creation of a source ontology is basically manually driven where incomplete information will be added.

The second phase is to extract concepts or entities and their relationships from a patient’s medical report using IE techniques which had been customized in the first phase. The extracted data will be presented in the form of an ontology and stored as a target ontology.

The third phase is to create an ontology manager. The purpose of the ontology manager is to map between the target ontology and the source ontology and to produce a mapping ontology. The mapping ontology will be presented to a user (clinician) in a visualization form using scheme colours to enhance the readability and understanding.

Visually representing the content of a text document is one of the most important operations in the field of text visualization. Specifically, visualization techniques have been developed to show the content of documents from different aspects and at different levels of details [36]. Applying visual representation in knowledge discovery aims at facilitating the understanding of its results. Thus, visualization techniques could be integrated into a process of recognizing and understanding extracted medical data from patient’s medical history reports. The process is useful for viewing the data to be analysed and understanding the results of extracted data [37]. Colour mapping has been seen as an important technique used in visualization to build easy to discern visual representations of data and information [38]. Colour selection in data visualization is not only as an aesthetic choice, but also it is a crucial tool to convey the meaning of the extracted data accurately. Differences in colour can distinguish different groups of concepts [39]. The use of colour in COBIE will help users to capture the presented knowledge with less effort.

4. Conclusion
We have, in this paper, touched on some aspects of medical information extraction research works and we have proposed a theoretical framework on ontology-based clinical information extraction. The proposed system involves a combination of IE and creating ontology for facilitating the process of extracting and visualizing clinical information. Despite the fact that there is a reasonable of research work that has been reported in this field, extracting medical concepts from medical reports is still challenging for researchers. This paper serves as a starting point for researchers to utilize IE and ontology for medical information extraction.

References
[1] Sun Y and Loparo K 2019 Information Extraction from Free Text in Clinical Trials with Knowledge-Based Distant Supervision In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) vol 1 pp 954-955
[2] Jurafsky D 2000 Speech & language processing Pearson Education India
[3] Jusoh S and Alfawareh HM 2012 Techniques, applications and challenging issue in text mining International Journal of Computer Science Issues (IJCNI) 9(6)
[4] Hobbs JR and Riloff E 2010 Information Extraction. Handbook of natural language processing
[5] Allahyari M Pouriyeh S Assefi M Safaei S Trippe ED Gutierrez JB and Kochut K 2017 A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.
[6] Dou D Wang H and Liu H 2015 February. Semantic data mining: A survey of ontology-based approaches. In Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015) (pp. 244-251). IEEE.
[7] Sugumaran V 2016 Semantic technologies for enhancing knowledge management systems. In Successes and Failures of Knowledge Management pp 203-213 Morgan Kaufmann
Wimalasuriya D C & Dou D 2010 Ontology-based information extraction: An introduction and a survey of current approaches Journal of Information Science 36(3) pp 306–323 https://doi.org/10.1177/0165551509360123

Khin NPP, Lynn KT. 2017 Medical concept extraction: A comparison of statistical and semantic methods. In: 2017 18th IEEE/ACIS International Conference on Software Engineering, Articial Intelligence, Networking and Parallel/Distributed Computing (SNPD) p 35-38.

Jurafsky D, Martin JH 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR;

Yu H, Hripcsak G, Friedman C. 2002 Mapping abbreviations to full forms in biomedical articles. Journal of the American Medical Informatics Association. 9(3): pp 262-272.

Amer E, Fouad KM. 2016 Keyphrase extraction methodology from short abstracts of medical documents. In: 2016 8th Cairo International Biomedical Engineering Conference (CIBEC) p. 23-26

Patrick J, Min Li. 2009 Intelligent Clinical Notes System: An information retrieval and information extraction system for Clinical Notes. In: 2009 11th International Conference on e-Health Networking, Applications and Services (Healthcom) p. 108-115.

Lafferty J, McCallum A, Pereira FC 2001 Conditional random fields: Probabilistic models segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), p 282-289

McCallum A, Freitag D, Pereira FC. 2000 Maximum Entropy Markov Models for Information Extraction and Segmentation. In: Proceeding 17th International Conf. on Machine Learning p 591-598

Xu R Supekar K Huang Y Das A and Garber A 2006 Combining text classification and hidden markov modeling techniques for structuring randomized clinical trial abstracts. In AMIA Annual Symposium Proceedings vol p. 824 American Medical Informatics Association

Chen X Shi S Zhan S Jiang D Lin X 2019 Named Entity Recognition of Chinese Electronic Medical Records Based on Cascaded Conditional Random Field. In: 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) p 364-368.

Aronson AR 200 Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program In: Proceedings of the AMIA Symposium American Medical Informatics Association p 17

He Z Weng H Ou A Yan S Lu C Li G 2017 Feature extraction from medical record text for TCM classification of psoriasis. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM p 1354-1356

Bhaskoro SB, Supangkat SH 2014 An extraction of medical information based on human handwritings. In: 2014 International Conference on Information Technology Systems and Innovation (ICITSI) p 253-258

Jin Z 2018 Ontology-Oriented Interactive Environment Modeling. In: Environment Modeling-Based Requirements Engineering for Software Intensive Systems Elsevier p 45-67 Available from: https://doi.org/10.1016/b978-0-12-801954-2.00004-2

Charniak E 2001 Unsupervised learning of name structure from coreference data. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies Association for Computational Linguistics p. 1-7

Bloehdorn S Cimiano P Hotho A Staab S 2005 An Ontology-based Framework for Text Mining. In: LDV Forum vol 20 p 87-112

Mala V Lobiyal DK 2015 Concepts extraction for medical documents using ontology In: 2015
International Conference on Advances in Computer Engineering and Applications p 773-777

[25] Shen Y Li Y Deng Y Zhang J Yang M Chen J Si S and Lei K 2018 Gastroenterology ontology construction using synonym identification and relation extraction IEEE Access 6 pp 52095-52104

[26] Le Moigno S Charlet J Bourigault D Degoulet P Jaulent MC 2002 Terminology extraction from text to build an ontology in surgical intensive care. In: Proceedings of the AMIA Symposium. American Medical Informatics Association p 430

[27] Mate S Köpcke F Toddenroth D Martin M Prokosch HU Bürkle T and Gansland T 2015 Ontology-based data integration between clinical and research systems PloS one 10(1) p.e0116656

[28] Mannai M Ben Abdessalem Karaa W 2013 Bayesian information extraction network for Medline abstract In: 2013 World Congress on Computer and Information Technology (WCCIT) p. 1-3.

[29] Baneyx A Charlet J Jaulent MC 2006 Methodology to build medical ontology from textual resources In: AMIA Annual Symposium proceedings American Medical Informatics Association p 21

[30] Zhu W Fu L Xu L Zhang B A TCM 2011 Diagnosis System Based on Textbook Information Extraction In:2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing p 483-487

[31] Sahay S Li B Garcia EV Agichtein E Ram A 2007 Domain Ontology Construction from Biomedical Text International Conference on Artificial Intelligence (ICAI’07) CSREA Press p. 28-34

[32] Alamu F Aworinde H Oparah O 2014 Development of a semantic ontology for Malaria disease using Protege - OWL software Journal of Multidisciplinary Engineering Science and Technology (JMEST) 1(3)

[33] Jiang L Zhang HB Yang X Xie N 2012 Research on semantic text mining based on domain ontology In: International Conference on Computer and Computing Technologies in Agriculture Springer p 336-343

[34] Girardi D Dirnberger J Giretzlehner M 2015 An ontology-based clinical data warehouse for scientific research Safety in Health 1(1):6

[35] Alani H 2006 Ontology Construction from Online Ontologies In proc. In 15th International World Wide Web Conference

[36] Cao N and Cui W 2016 Overview of text visualization techniques In Introduction to Text Visualization (pp. 11-40) Atlantis Press Paris

[37] Dias MM Yamaguchi JK Rabelo E Franco C 2012 Visualization Techniques Which is the Most Appropriate in the Process of Knowledge Discovery in Data Base? Advances in Data Mining Knowledge Discovery and Applications p 155-177

[38] Silva S Santos BS Madeira J 2011 Using color in visualization A survey Computers & Graphics 35(2) p 320-333

[39] Pajevic S Pierpaoli C 2000 Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: application to white matter fiber tract mapping in the human brain Magnetic resonance in medicine. 43(6) p 921-921