EUROPEAN ACADEMIC RESEARCH
Vol. VIII, Issue 12/ March 2021

Impact Factor: 3.4546 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

Comparison of medical ontologies and their development methodology

PETRIKA MANIKA
Department of Informatics, Faculty of Natural Sciences
University of Tirana, Albania
ANA KTONA
Department of Informatics, Faculty of Natural Sciences
University of Tirana, Albania

Abstract

Ontology refers to the state of being and is used to describe entities and concepts of a domain. Ontologies are widely applied in medicine and describe medical concepts, terminologies and the relationships between them. Ontologies can be used along with the semantic web to share the medical information. This paper aims to highlight some of the most popular ontologies in the field of medicine concerning computer science discipline. Identifying flawed practices and anomalies in ontology is an important subject to be addressed by researchers. This paper gives an introduction to ontologies in the field of medicine and some of its purposes. NLP (Natural Language Processing) is important in different computer science fields like information extraction, machine translation, question answering, etc. NLP can be used to help create medical ontologies. This paper also discusses the different medical ontologies, the methods and approaches for each and also an architecture for software development based on medical ontologies and conclusions.

Keywords: Natural Language Processing, ontology, semantic web, lexico-syntactic pattern, text mining

INTRODUCTION

Ontology comes from the word onto-logos and is the science of being. Due to the fragility of the subject, there are many debates on which

entities exist in medicine. These entities include biological structures, substructures, pathogenic agents, diseases, and therapeutic effects, among others. In medicine, ontology is the study that aims to ascertain the entities that exist in the world of medicine, and the relationships between them (Maynard, Bontcheva & Augenstein 2016, p. 1). Medical ontology is needed to avoid confusion and determine entities that are presupposed as true in medicine. Usually, all structured medical terminologies are referred to as medical ontologies. There is a variety of medical ontologies for different medical domains like genetics, anatomy, drugs, adverse events, diseases and medical codes. The NCBO BioPortal is a library that has over 500 medical ontologies and terminologies covering nearly 8 million medical concepts (Bail et al. 2011, p. 67). Medical ontologies seek to provide people with a standard representation and vocabulary for describing and data analysis to derive meaningful inferences. Medical ontologies are built to facilitate the sharing of knowledge among researchers. Some of the main purposes of ontology was sharing knowledge on the same domain, reuse of previously used ontology, and sharing domain knowledge from operational knowledge. Doctors and healthcare scientist have their special medical language that they use to effectively communicate general medical knowledge and patient information. Nowadays, hospitals collect large volumes of data from patients which is referred to as medical big data. Big data refers to a large volume of data, either structured and unstructured, that inundate the day to day activities of a business. Ontologies have been created in the medical field to represent this data (Gai et al. 2015, p. 752). Ontologies are essential in the scope of recovery, organization of information and the semantic web. Recently ontologies have been used in the field of medicine in the area of natural language processing. This paper covers a methodology used to extract knowledge from data stored about patients' clinical records and general medicine to realize relevant knowledge elements for researchers. This paper aims to analyze a methodology that allows for text written in natural language in a computer to be translated using natural language processing tools to give out essential information to medical specialists. Comparative methodology has been used to highlight the most popular ontologies in the field of medicine.

SEMANTIC WEB

The semantic web extends the World Wide Web (WWW) or the web, whose aims to deliver a standard framework for sharing and reusing data across the Internet (Panahiazar et al. 2014, p. 790). It ultimately wants to structure machines to enable machines to interpret information on the current web without human intervention. Semantic web uses agent technology, ontology, and some standard mark-up languages, such as RDF (Resource Description Framework), and OWL (Ontology Web Language) to model information from web resources. Ontology makes it easier to access, find, maintain or present electronic information that is available on the Internet. Thus, the semantic web can be used to improve the keyword search problem. It allows a user to perform queries that are semantic-based and search for the information they require. Ontologies improve the semantics by providing richer relationships between the terms of a vocabulary. Ontologies are logic-based which means that detailed, and meaningful inferences can be made between data. Semantic web ontology can improve communication between humans and computers. The major uses of sematic web ontology are to help improve in communication between humans, to provide interoperability and allow communication and to optimize the design and the final quality of software systems.

Sharing of knowledge and semantic interoperability among bio-medical information systems become more easy with semantic web technologies. Semantic web can help facilitate the process of extracting knowledge from the heterogeneous data and knowledge sources in healthcare. Decision making in the healthcare field requires information sharing, so ontologies helps the clinicians to collect the right information and avoid repeating the experiences. Ontologies add context to the patient's medical history and thanks to the definition of relationships, ontologies create links among diagnosis and medications, laboratory tests and radiology examination automatically. This approach makes queries more effective and the results are closer to search terms. The semantic web provides a common framework for sharing and reuse of knowledge among systems and organizations. This sharing and reuse of knowledge improves scientific research through creating new ideas, testing

different hypotheses from different aspects, facilitating the training of novice researchers and reduce the costs of information gathering.

MEDICAL ONTOLOGIES

The applicability of ontology in the field of healthcare and medicine is no longer just a research topic. There are many groups of scientists and universities that build and manage medical ontologies. Oncological medical research program, is a program developed by Lister Hill National Center for Biomedical Communications (part of the National Library of American Medicine). The purpose of this program is to develop a medical ontology which enables various applications to process knowledge in order to communicate easily between them. Cooperative open ontology development environment is a project developed by the Medical Informatics Group at the University of Manchester. This project aims to provide support to communities interested in OWL by developing materials, raising tools, and exploring some of the theoretical problems to help usable solutions before major problems arise (CO-ODE project, n.d.). Below are some of the concrete cases where ontologies in the field of medicine are applied.

LinKBase - is a biomedical ontology. It is one of the best ontologies for supporting NLP / NLU (Natural Language Processing and Natural Language Understanding) and data integration applications, due to its hierarchical structure, coverage, use of operational, formal and linguistic relationships, combined with its basic language technology. LinKBase is designed with the primary goal of integrating terminology and databases applications designed for NLP and information management and retrieval. This biomedical ontology covers various aspects of medicine, including procedures, anatomy, pharmaceuticals, and various disorders and anomalies that provide over 9 million elements of knowledge, making it the largest biomedical knowledge base in the world (van Gurp et al. 2006).

ONIONS - (Ontologic Integration on Naïve Sources) It is a methodology that creates a common framework for interpreting the definitions used to organize a range of terminological sources. In other words, it allows to coherently process a field ontology for each resource, which can then be compared to others and defined in an

integrated model. It is designed to build the ON9 medical ontology (Gangemi et al. 2000).

GALEN - (Generalized Architecture for Languages, Encyclopaedias, and Nomenclatures in medicine), is a European Union project that aims to provide reusable terminological resources for clinical systems. This ontology aims to represent all medical concepts including sensitive concepts regardless of any application. A key feature of GALEN is that it was built form a predefined set of high-level knowledge representations before populating the ontology. Unlike traditional terminological resources, which are pre-coordinated, GALEN essentially provides the building blocks required to describe terminologies. GALEN ontology provides a mechanism for combining simple concepts. An open foundation was established in 2000 under the name OpenGALEN which contains about 25,000 concepts. This foundation was established to share the reference model and work with software vendors and terminology developers to support its expansion and use. Since then the Galen model has been used to study healthcare terminology, surgical procedure, and anatomy (Bodenreider & Burgun 2005). Table shows the metrics of GALEN ontology (GALEN ontology 2017).

Classes	23141
Individuals	0
Properties	950
Maximum depth	24
Maximum number of children	1492
Average number of children	4
Classes with a single child	1944
Classes with more than 25 children	120
Classes with no definition	21974

Tab 1. Metrics of GALEN ontology

UMLS® - The Unified Medical Language System® was developed by the National Library of Medicine to help healthcare professionals and researchers to access biomedical information from a variety of sources. The Metathesaurus is the biggest component of the UMLS. It is a large biomedical thesaurus that is organized by concept, or meaning, and it links similar names for the same concept from nearly 200 different vocabularies. It integrates over a million concepts from more than a hundred dictionaries and terminology. While the structure of

each source is preserved in the construction of Metathesaurus, the equivalent terms are grouped into a unique semantic concept. Concept relationships are either inherited from the basic vocabulary or generated specifically. Since Metathesaurus does not have resource constraints, it cannot provide the type of organization expected from an ontology. In contrast, the Semantic Web was developed independently of the dictionary integrated into Metathesaurus and serves as a basic high-level ontology for the biomedical field (Bodenreider & Burgun 2005).

GO - Gene Ontology project provides the most comprehensive resource available for computable knowledge about gene functions and products. The GO knowledge base consists of two main components (Gene ontology, n.d.):

- GO provides the logical structure of biological functions and their relationship to each other, displaying it as an acyclic directed graph.
- The corpus of GO notes and evidence-based statements are related to a specific genetic product in a particular ontological term.

The authors found it reasonable to put definitions and comments on genes as notes in ontologies. The most common use of Gene ontology records is for interpreting large-scale molecular biology experiments, sometimes called "omics" experiments, which measure:

1) Genetic products (RNA and proteins).
2) Variations on the DNA gene sequence.
3) Small molecules metabolized by proteins.

The following table shows the GO metric (GO ontology 2021):

Classes	49290
Individuals	0
Properties	9
Maximum depth	16
Maximum number of children	9485
Average number of children	4
Classes with a single child	5054
Classes with more than 25 children	393
Classes with no definition	2170

Tab 2. Metrics of GO ontology

CPR - Computer-based Patient Record according to the IOM (Institute of Medicine) is an electronic patient register that exists in a system designed specifically to support users, providing access to structured data, clinical decision support systems, links to medical knowledge and other equipment (Ogbuji 2011). In its current form, this ontology is a high-level framework for the collections of clinical vocabulary systems, where all concepts are given a canonical and semantic syntax. It has been used as a coordination system for managing or deriving precise meanings of terminology used by healthcare professionals as well as researchers for use in applications involving large-scale data management. Like many contemporary ontologies that have a related purpose, CPR ontology is based on basic formal ontology (CPR ontology 2013).

Classes	128
Individuals	32
Properties	38
Maximum depth	8
Maximum number of children	9
Average number of children	2
Classes with a single child	10
Classes with more than 25 children	0
Classes with no definition	33

Tab 3. Metrics of CPR ontology

RiboWeb - is a resource that facilitates the construction of three-dimensional models of ribosomal components. RiboWeb is used also to compare the results to existing studies. Parts of it are four ontologies that contain the knowledge used to perform these tasks. The ontologies of RiboWeb are the physical-thing ontology, the data ontology, the publication ontology and the methods ontology. The physical-thing ontology is used for the physical side of the domain, to describe ribosomal components and associated `physical things'. The data ontology is used as a knowledge base for experimental details and for the structure of physical-things. The methods ontology contains information about techniques for analyzing data. It is used as a knowledge base of which techniques can be applied to which data, and also as a knowledge for the input and outputs of each method. The instances that are added to RiboWeb correspond to these concepts (Chen et al. 1997).

NATURAL LANGUAGE PROCESSING

NLP is related to human-computer interaction. Understanding NLP is important in sectors like information extraction, machine translation and question answering, among others (Simon et al. 2006, p. 224). Because of this, the production of software tools and the semantic web has risen. These tools are mostly available for free on the Internet, but unfortunately, there is a language barrier as most only work with known languages like English, French and Spanish. Therefore, there is a need to represent other natural languages using translations by NLP tools.

In natural language processing, three things are done, which include understanding the natural language, generating natural language and translation. There are several approaches of NLP to ontology learning, including symbolic, statistical and hybrid approaches. The symbolic approach utilizes information from linguistics to extract information from text. Conceptual relationships between terms can be identified using linguistic rules within an ontology. One example of a symbolic approach is to use the lexico-syntactic pattern (LSP) matching (Gupta et al. 2014, p. 902). In this approach, concepts are often represented as multi-term words, compound in nature, which in general are more specific than single compositional terms. This method assumes that a compound term is a hyponym of a single term. The statistical approach classifies words based on their meaning as well as on their co-occurrence with other words. This method has an advantage of little prior knowledge of a word is needed. Statistical methods are categorized into clustering and machine learning methods.

Ontologies can facilitate language processing in two main ways. When building the lexicon, an ontology can be used directly to define the terms, their concepts and relations for content word. Ontology is defined as a knowledge base and as such is expressed in a formal language. Ontologies provide knowledge for complex language processing. This knowledge if formally expressed as a structured list of concepts, relations and individuals. The ontology provides definitions for these, through the taxonomy relation between the terms and the properties specified for them. Ontologies are closely connected to NLP. Software adapted for building domain ontologies of natural languages can be found on the internet, but unfortunately

they cannot work for any given natural language. A methodology and NLP tool that enables one to create, visualize and manipulate ontologies in different representation formats is chosen for the architecture. The architecture created follows the functional requirements for the system and is used to develop the project according to the standards of other software in the market. There are free software tools on the Internet that can be used to perform analysis, design and implementation of computer programs with NLP capabilities. At all times, good practices of software development should be followed, and appropriate design patterns should be used, which will facilitate adaptability to changes and maintenance of the tool. The developed architecture will help the progress of the construction of NLP automatic tools that can carry out the processes of construction and population of ontologies. Architectures and NLP tool systems require a large number of clinical records and information to be successfully implemented.

CONCLUSIONS

Healthcare is a field where exists many terms, concepts and relations between them. Standardizing the definition of terms and the connections between them in a uniquely structured format is difficult and this puts barriers on system interoperability and integration with other healthcare system. A large volumes of data from patients and clinical records can be collected from hospitals. Building knowledge base from this data that can be processed by computers and shared on the internet is not an easy process. The semantic web and its technologies can be used for searching from knowledge bases on the internet and for sharing them. One of the most used semantic network technologies is ontology. Ontologies describes medical concepts, terminologies and the relationships between them. Ontologies can be used along with the semantic web to share the medical information. Nowadays there are many ontologies built in the field of medicine which can be accessed online and can be used by doctors and scientists all over the world. This paper discussed the different medical ontologies and the methods and approaches for each. Ontologies can be developed from big data that came from hospital records. The use of NLP techniques is required to process the data and to help create medical ontologies. NLP tools are mostly available

for free on the internet, but unfortunately, there is a language barrier as most only work with known languages. Using translations by NLP tools can help to remove this barrier. This paper also introduced an architecture for software development based on medical ontologies and conclusions.

REFERENCES

1. Bail, S, Horridge, M, Parsia, B & Sattler, U 2011, 'The justificatory structure of the NCBO bioportal ontologies', *International Semantic Web Conference*, pp. 67-82, https://link.springer.com/chapter/10.1007/978-3-642-25073-6_5.
2. Gai, K, Qiu, M, Chen, L-C & Liu, M 2015, 'Electronic health record error prevention approach using ontology in big data', in *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, pp. 752-7, https://ieeexplore.ieee.org/abstract/document/7336248/.
3. Gupta, S, MacLean, DL, Heer, J & Manning, CD 2014, 'Induced lexico-syntactic patterns improve information extraction from online medical forums', *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 902-9, <https://academic.oup.com/jamia/article-abstract/21/5/902/761575>, https://academic.oup.com/jamia/article-abstract/21/5/902/761575.
4. Maynard, D, Bontcheva, K & Augenstein, I 2016, 'Natural language processing for the semantic web', *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 6, no. 2, pp. 1-194, https://www.morganclaypool.com/doi/abs/10.2200/S00741ED1V01Y201611WBE015, https://www.morganclaypool.com/doi/abs/10.2200/S00741ED1V01Y201611WBE015.
5. Panahiazar, M, Taslimitehrani, V, Jadhav, A & Pathak, J 2014, 'Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases', *2014 IEEE International Conference on Big Data (Big Data)*, pp. 790-5, https://ieeexplore.ieee.org/abstract/document/7004307/.
6. Simon, J, Dos Santos, M, Fielding, J & Smith, B 2006, 'Formal ontology for natural language processing and the integration of biomedical databases', *International journal of medical informatics*, vol. 75, no. 3-4, pp. 224-31, <https://www.sciencedirect.com/science/article/pii/S1386505605001309>, https://www.sciencedirect.com/science/article/pii/S1386505605001309.
7. CO-ODE project. n.d. Accessed February 13, 2021. http://owl.cs.manchester.ac.uk/research/co-ode/.
8. Van Gurp M, Decoene M, Holvoet M & Casella dos Santos M, 'LinKBase, a Philosophically-inspired Ontology for NLP/NLU Applications', *CEUR Workshop Proceedings*. 222.

9. CO-ODE project. n.d. Accessed February 13, 2021. http://owl.cs.manchester.ac.uk/research/co-ode/.
10. Gangemi, A, Steve, G & Giacomelli, F. 2000, 'ONIONS: An Ontological Methodology for Taxonomic Knowledge Integration', https://www.researchgate.net/publication/2386179_ONIONS_An_Ontological_Methodology_for_Taxonomic_Knowledge_Integration
11. Bodenreider O & Burgun A 2005, 'Medical informatics: Advances in knowledge management and data mining in biomedicine', *Springer-Verlag*; 2005
12. GALEN ontology. 2017. Last modified January 16, 2017. https://bioportal.bioontology.org/ontologies/GALEN.
13. Bodenreider O & Burgun A 2005, 'Medical informatics: Advances in knowledge management and data mining in biomedicine", Springer-Verlag; 2005
14. Gene Ontology. n.d. Accessed February 19, 2021. http://geneontology.org/docs/introduction-to-go-resource/.
15. GO ontology. 2021. Last modified February 3, 2021. https://bioportal.bioontology.org/ontologies/GO.
16. Ogbuji, C 2011, 'A Framework Ontology for Computer-Based Patient Record Systems.', *ICBO* (2011).
17. CPRO ontology. 2013. Last modified April 24, 2013.https://bioportal.bioontology.org/ontologies/CPRO
18. Chen, R, Felciano, R & Altman, R. 1997, 'RIBOWEB: Linking Structural Computations to a Knowledge Base of Published Experimental Data.', *International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*. 5. 84-7.