ORIGINAL PAPER

An Ontology-Based Knowledge Methodology in the Medical Domain in the Latin America: the Study Case of Republic of Panama

Denis Cedeno-Moreno[1],
Miguel Vargas-Lombardo[2]

[1,2]Technological University of Panama, Research Group Electronics Health and Supercomputing, Panama City, Panama

Corresponding author: Miguel Vargas-Lombardo, Eng, MCs, PhD. Electronics Health and Supercomputing Research Group, Technological University of Panama, Panama city, Panama. ORCI https://orcid.org/0000-0002-2074-2939. E-mail miguel.vargas@utp.ac.pa

ABSTRACT

Introduction: Nowadays in Panama, there is a lot of patient information stored in textual form which cannot be manipulated to manage adequate knowledge. There are multiple resources created to represent knowledge, including specialized glossaries, ontologies, among others. The ontologies are an important part within the scope of the recovery and organization of the information and the semantic web. Also in recent works they are used in applications of natural language processing (NLP), as a knowledge base. **Aim:** This research was conducted with the aim of creating a methodology that allows from a text written in NL, extract the necessary elements using NLP tools and with them create a knowledge base represented by one domain ontology and extract knowledge to help medical specialists. **Material and Methods:** In this study we carried out a methodology that allows the extraction of knowledge of patient clinical records, general medicine and palliative care, in order to show relevant knowledge elements to specialists. The methodology was validated with a data corpus of approximately 200 patient records. **Conclusion:** We have created a knowledge representation methodology, combining NLP techniques and tools and the automatic instantiation of an ontology, which can serve as a software agent for other applications or used to visualize the patient's clinical information. The study was validated using the traditional metrics of information retrieval systems precision, recall, F-measure obtaining excellent results, and can be used as a software agent or methodology for the development of information extraction software systems in the medical domain.

Keywords: Natural language processing, knowledge base, information extraction, information retrieval.

1. INTRODUCTION.

The convergence of different areas of knowledge has led to the design and implementation of computer systems that support the integration of innovative tools. Medical knowledge, represented in medical information, is most of the time in textual documents, written in NL, that is, in unstructured data sources, to manage it, NLP techniques are used (1).

The first step for the computer processing of linguistic knowledge is the formal representation of such knowledge, there are multiple resources created to represent linguistic information, including specialized glossaries, taxonomies, thesauri and ontologies (2). The formal semantics underlying an ontology allows the automatic processing of information and allows the use of semantic reasoners to infer new knowledge. In this work we consider an ontology as "a formal and explicit specification of a shared conceptualization" (3).

Ontologies provide a formal and structured representation of knowledge and have the advantage of being reusable and can be shared. They also provide a vocabulary for a specific domain and define with different levels of formality the meaning of terms and the relationships between them. The knowledge in ontologies is formalized mainly using 5 types of components: classes, relationships, functions, axioms and instances (4).

Also, ontologies are an important part within the scope of the recovery and organization of information and the semantic web. In addition, they are becoming increasingly important within the NLP (5). Ontologies can be created manually, however this results in various cost and time problems
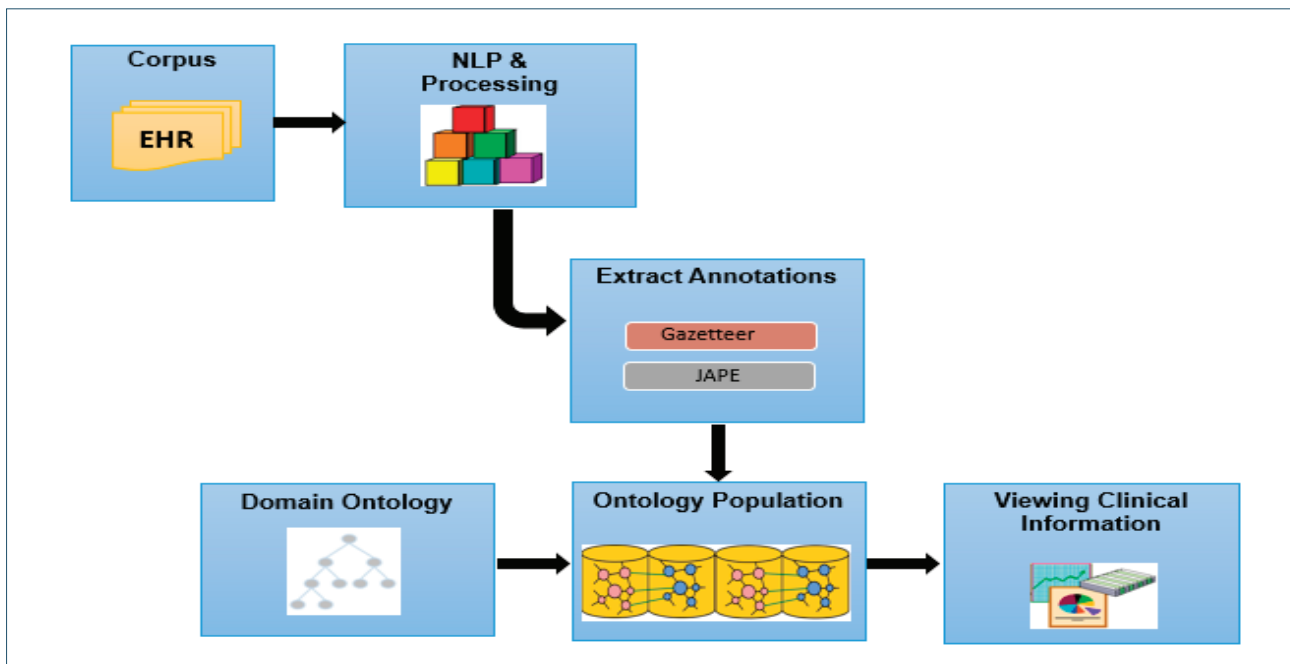
Figure 1. Phases of the proposed software architecture implemented within the primary care clinic of the Technological university of Panama.

(6). As an alternative, automatic learning of ontologies arises from textual documents whose objective is to identify the ontological elements automatically or semi automatically. It is an interesting approach that attempts to reduce time and resources. This is done using techniques and methods from fields such as artificial intelligence (AI), machine learning (ML), information retrieval (IR) or NLP.

This work proposes a methodology that allows from a text written in NL, extract the necessary elements using NLP tools to then instantiate an ontology automatically and extract knowledge. To then visualize it in a friendly way by the stakeholders through a conceptual map.

The structure of the document can be described as follows: section 2 show the aim of study; the section 3 materials and methods; section 4 explains the discussion of the work research on the software architecture and the conclusions are presented in section 5.

2. AIM

The objective of the present study is to provide an adequate methodology to allow the extraction of information from a clinical (7) text written in NL, which represents the corpus of patients medical records (8), extract the named entities and the relevant knowledge elements, and generate and create an instance of an ontology of the domain, which then serves as a software agent for other processes.

Methodology consists of four stages: (i) The stage of NLP and corpus processing, (ii) the stage extraction of annotations (iii) the population stage of the ontology and (iv) the stage of visualization of clinical information. In addition, the methodology must comply with semantic interoperability restrictions and, more specifically, it should be possible to extract all the information from patients in the HL7 (9) electronic clinical records standard. Figure 1 shows the phases of this architecture.

3. MATERIALS AND METHODS

To validate the methodology, we made a descriptive study in two hospitals of the Republic of Panama. In 2017, we obtained a corpus of clinical information of patients, specifically the specialties of palliative care and general medicine. The total corpus is made up of 200 documents, 100 from the palliative care domain (10) and 100 from general medicine. These documents contain information about the general patient, their attendants, clinical diagnoses, medication, history of laboratory services and other problems written in Spanish. This information is what the specialist assesses and records once a patient is seen in the consultation. Clinical information of patients used in the corpus were obtained from the primary care clinic of the Technological university of Panama.

The total corpus of the experiment contains approximately 98,000 words distributed throughout these 200 records of medical records. Finally, the methodology was constructed as a software architecture which was used and evaluated by the experts. It was evaluated using precision, recovery and F-measure validation metrics (11).

4. RESULTS AND DISCUSSION

We have decided to use the GATE tool that offers techniques and methods of ML (12) and we achieve IR (13) by programming a set of JAPE rules that extract and label patient information and then make the annotations defined in the rules. It was developed in Java and that gives it the advantage of being able to work under different operating systems. It also implements a broad set of knowledge modeling structures and actions that allow the creation, visualization and manipulation of ontologies in various representation formats.

Architecture that has been implemented, follows functional requirements and develops the project according to the standards required by the software industry. The technological foundations presented are those that have been considered necessary to create a software system that carries out the NLP from text in NL. It was decided to use free software tools that

are available to the research community, to perform the analysis, design and implementation of this computer program application. At all times good practices of software development have been followed and design patterns have been used, which facilitates adaptability to changes and maintenance of the tool. Figure 1 shows the phases of this architecture.

For the validation of the architecture we have used the clinical information domain of patients. The total corpus is made up of 200 documents. Furthermore, to achieve a better clarity in the validation of data, each corpus has been segmented into different sections: general data, diagnosis, medication.

Annotation extraction process was evaluated, using the precision, recall and F-measure validation metrics. We have obtained total experimental results for the document annotation extraction process, where we have the amount of general data annotations 2,764. To diagnose 505 extractions and 590 for the medication section. Table 1 below shows the total experimental results for the extraction of document annotations.

Type information	Quantity
General data	2,764
Diagnosis	505
Medication	590

Table 1. Total experimental results for the extraction of document annotations.

These values of general data, diagnosis, medication, we have established the evaluation criteria obtaining 95.5% of precision, 88.5% of recall and 91.9% of F-measure respectively. According to other studies and comments from experts and professionals that have been studied, the metrics (precision, recall and F-measure) (14, 15) implemented in the architecture include specific and general criteria for the evaluation of this type of information extraction methodologies. The comparison of these metrics specified in the methodology are objective and possible, so that the result of the evaluation is accurate.

Most of the medical information is currently in unstructured form, in textual documents (16, 17). There are some computer technologies that help us in this type of problem. In recent years, different approaches have been applied to extract knowledge from textual documents. Overall, in this study, we architecture allows that from a medical text written in NL, the necessary elements can be extracted using NLP tools to then instantiate an ontology automatically and extract knowledge.

It should be noted that to develop the components of the architecture, only free software technologies were used, which results in the fact that there is no need to pay for licenses for the development of the project. At present there are very few ontological learning systems oriented to the information domain of the patient for the construction of ontologies, therefore the investigations carried out in this field are increasingly important.

The contributions given in this architecture, help the progress of the construction of NLP automatic tools that can carry out the processes of construction and population of ontologies. It also provides an intermediate file in OWL (18, 19) format that can be consulted with an ontological editor such as Protégé, where you can see the results. A file is also generated in XML format where the HL7-CDA clinical information and message interoperability standard (20) is implemented. The interface of this software architecture has been designed so that it can be used by both NLP expert users as well as domain experts.

5. CONCLUSION

From the obtained results we can conclude the methodology presented in this study has been developed using a large amount of resources and documents of medical records of patients from the Republic of Panama. The objective is to support the processes of collection and extraction of knowledge in the medical domain, in areas such as palliative care as well as in general medicine.

One of the advantages of our methodology is that it was validated and approved by the experts. Therefore, software programmers can use it as a basis to design, implement and produce software in other domains. Our proposed methodology follows the practice of the use of free software, is simple and can be used in other domains, such as finance, education.

REFERENCES

1. Cimiano P, Unger C, McCrae JP. Ontology-Based Interpretation of Natural Language. Synth Lect Hum Lang Technol. 2014; 7(2): 1-178.
2. Danger R, Berlanga R. Generating complex ontology instances from documents. J Algorithms. 2009; 64(1): 16–30.
3. Studer R, Benjamins R, Fensel D. Knowledge Engineering: Principles and Methods. Data Know Eng. 1998; 25(1–2): 161–197.
4. Gruber TR. A translation approach to portable ontology specifications. Know Acquis. 1993; 5(2): 199–220.
5. Barrett N, Weber-Jahnke JH, Thai V. Engineering natural language processing solutions for structured information from clinical text: Extracting sentinel events from palliative care consult letters. Stud Health Technol Inform. 2013;192:594–598.
6. Kordjamshidi P, Moens M. Global Machine Learning for Spatial Ontology Population. J Web Semant. 2015; 30: 3–21.
7. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C. Electronic health records: New opportunities for clinical research. J Intern Med. 2013; 274(6): 547–560.
8. West SL, Johnson W, Visscher W, Kluckman M, Qin Y, Larsen A. The challenges of linking health insurer claims with electronic medical records. Health Informatics J. 2014; 20(1): 22–34.
9. Del Fiol G, Huser V, Strasberg HR, Maviglia SM, Curtis C, Cimino JJ. Implementations of the HL7 Context-Aware Knowledge Retrieval (" Infobutton ") Standard: Challenges, strengths, limitations, and uptake. J Biomed Inform. 2012; 45(4): 726–735.

10. Smith TJ, Coyne P, Cassel B, Penberthy L, Hopson A, Hager MA. A high-volume specialist palliative care unit and team may reduce in-hospital end-of-life care costs. J Palliat Med. 2003; 6(5): 699–705.

11. Quesada-Martínez M, Mikroyannidi E, Fernández-Breis JT, Stevens R. Approaching the axiomatic enrichment of the Gene Ontology from a lexical perspective. Artif Intell Med. 2015; 65(1): 35–48.

12. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: Review of current status and future directions. Int J Med Inform. 2014; 83(9): 605–623.

13. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. 2008; 9(2): S8.

14. Al-Nazer A, Helmy T, Al-Mulhem M. User's profile ontology-based semantic framework for personalized food and nutrition recommendation. Procedia Comput Sci. 2014; 32: 101–108.

15. Wong W, Liu WEI, Bennamoun M. Ontology learning from text : A look back and into the future. ACM Comput Surv. 2012; 44(4): 1–36.

16. Pande VC, Khandelwal AS. A Survey Of Different Text Mining Techniques. J Manag Res. 2014; 3(1): 125–133.

17. Fan H, Li H. Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques. Autom Constr. 2013; 34: 85–91.

18. Ochoa JL, Valencia-García R, Perez-Soltero A, Barceló-Valenzuela M. A semantic role labelling-based framework for learning ontologies from Spanish documents. Expert Syst Appl. 2013; 40(6): 2058–2068.

19. Zhou S, Ling H, Han M, Zhang H. Ontology Generator from Relational Database Based on Jena. Comput Inf Sci. 2010; 3(2): 263–267.

20. Malik SK, Prakash N, Rizvi S a M. Ontology Design and Development: Some aspects: An overview. Semant Web J. 2010;1–6.