Case Report

Healthc Inform Res. 2018 October;24(4):376-380.
https://doi.org/10.4258/hir.2018.24.4.376
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Design and Construction of a NLP Based Knowledge Extraction Methodology in the Medical Domain Applied to Clinical Information

Denis Cedeño Moreno, PhD, Miguel Vargas-Lombardo, PhD
Technological University of Panama, Panama City, Panama

Objectives: This research presents the design and development of a software architecture using natural language processing tools and the use of an ontology of knowledge as a knowledge base. **Methods:** The software extracts, manages and represents the knowledge of a text in natural language. A corpus of more than 200 medical domain documents from the general medicine and palliative care areas was validated, demonstrating relevant knowledge elements for physicians. **Results:** Indicators for precision, recall and F-measure were applied. An ontology was created called the knowledge elements of the medical domain to manipulate patient information, which can be read or accessed from any other software platform. **Conclusions:** The developed software architecture extracts the medical knowledge of the clinical histories of patients from two different corpora. The architecture was validated using the metrics of information extraction systems.

Keywords: Knowledge, Knowledge Management, Natural Language Processing, Information Extraction

I. Introduction

The innovations in areas such as health information technology (HIT) [1] have the potential of improving people's health and leading to better quality in modern health systems. Moreover, human knowledge becomes a powerful strategic tool for a health organization [2]. Having this knowledge depends on the ability of individuals performing certain tasks. Because ontologies [3] constitute the standard knowledge representation mechanism for the semantic web and other information retrieval systems, it becomes necessary to develop ontologies that are capable of depicting health information. A well-accepted definition in the area of artificial intelligence (AI) is that of Studer et al. [4], who said: "an ontology is a formal and explicit specification of a shared conceptualization."

The use of ontologies is becoming increasingly important in natural language processing (NLP) [5,6]. NLP is the discipline that deals with the automatic treatment of natural language [7]. It is a branch of AI and computational linguistics that is dedicated to understanding human language to exploit the linguistic knowledge of texts [8].

Submitted: July 25, 2018
Revised: 1st, July 30, 2018; 2nd, September 24, 2018; 3rd, October 26, 2018
Accepted: October 26, 2018

Corresponding Author
Miguel Vargas-Lombardo, PhD
Technological University of Panama, Via Centenario, Ancon, Panama. Tel: +50763005543, E-mail: miguel.vargas@utp.ac.pa (https://orcid.org/0000-0002-2074-2939)

II. Case Description

This work developed a software architecture that enables, from a text written in natural language, the extraction of the necessary elements using NLP tools to then automatically create an instance of an ontology and to extract medical knowledge from the patient records of Panamanians. The software architecture was tested and validated by experts using precision, recall and F-measure metrics, obtaining excellent results.

1. Proposed Methodology

We developed a software architecture that enables the extraction of information from a clinical text written in natural language, which represents the corpus of the patient clinical records, followed by the extraction of the named entities and relevant knowledge elements, and finally, the generation and the instantiation of an ontology of the domain. It has four stages: (1) NLP, (2) extraction of annotations, (3) population of the ontology, and (4) showing the clinical knowledge. The proposed architecture complies with semantic interoperability standards because it extracts patient information using the HL7 electronic clinical records standard. Figure 1 shows the phases of this architecture.

2. Corpus

The corpus consists of clinical data of about 200 patients, whose medical notes were obtained from the primary care clinic of the Technological University of Panama and a public hospital, both in Panama City. They contain general data, clinical diagnoses, medications, history of laboratory services and clinical aspects described in Spanish by physicians. In addition, the software architecture was evaluated by experts using precision metrics, recovery and F-measure, which are typical of information retrieval systems [9,10].

3. Natural Language Processing

In this step, the objective was to perform a linguistical analysis of the text. This is done by dividing the text into sentences and words. The standard word segmentation task was completed with the application of a programming interface provided in the development framework for the NLP called GATE (General Architecture for Text Engineering) [11,12].

4. Extraction of Annotations

For the extraction of information for labeling annotations, two GATE components called the Java Annotation Pattern Engine (JAPE) Transducer [13] and Gazetteer [14] were used. These components are responsible for compiling and executing a set of rules based on the JAPE grammar. A set of 65 semantic rules were constructed and coded in the JAPE grammar to extract the necessary annotations in this architecture. Figure 2 shows an extraction rule written in JAPE.

5. Ontology Population

In this phase, the instances that populate the ontology will be inserted. In the defined architecture, the annotations recovered in the previous phase are used to carry out the process of instantiating the ontology. In Figure 3, with the Protégé [15] editor, the class hierarchy of the created domain ontology is shown.

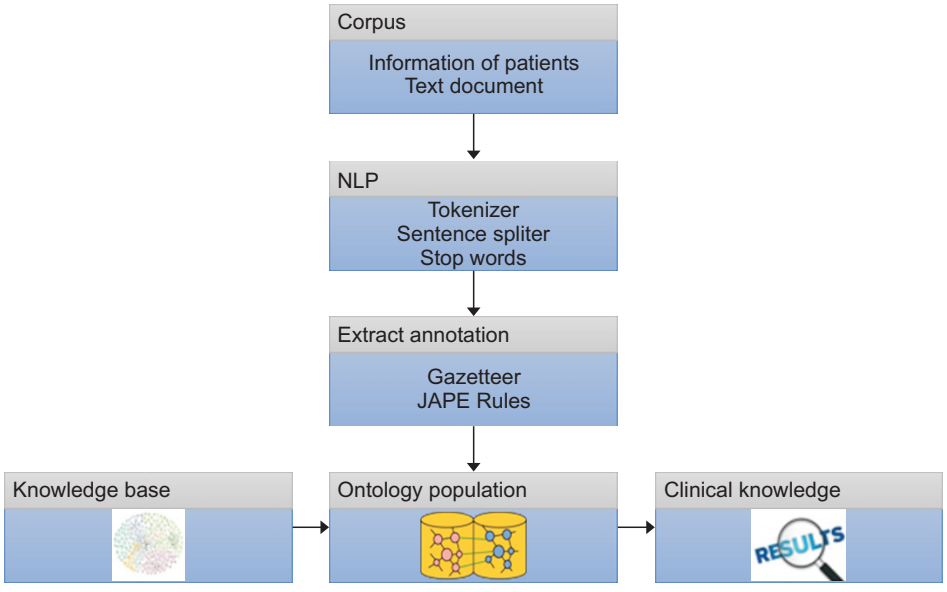

Figure 1. Phases of the proposed architecture. NLP: natural language processing, JAPE: Java Annotation Pattern Engine.

6. Extracting Clinical Knowledge

The result of the architecture expressed in clinical knowledge was, first, a file in OWL format [16], that can be read from any ontology editor, either to create software agents or to reuse it, enriching it with more ontological elements. Second, a friendly user interface is used for the health professional. It enables knowledge management through the extraction process that will be used as support for medical decision making. Third, an XML file is generated following the HL7 Clinical Document Architecture (CDA) standard [17] that presents patient information schematically and can be used by other systems which enables interoperability through this standard.

```
Phase: procedencia
Input: Token Lookup Split
Options: control = brill

Rule: procedenciaRule
Priority: 55
( {Token.string == "Procedente"})
( {Token.string == "de"})
( {Token.string == ":"})?
(( {Token})*):label
( {Split})

-->:label.ProcedenteDe = {rule =  "procedenciaRule"}
```

Figure 2. Excerpt of a rule code written in JAPE (Java Annotation Pattern Engine) for the extraction of annotations.

III. Discussion

Most health centers in Panama handle unstructured information. Physicians and nurses write a patient's history in a text document. This makes it difficult for computers to read and understand this information. To solve this problem, a methodology and a software architecture that makes it possible, using NLP techniques to automatically create an on-

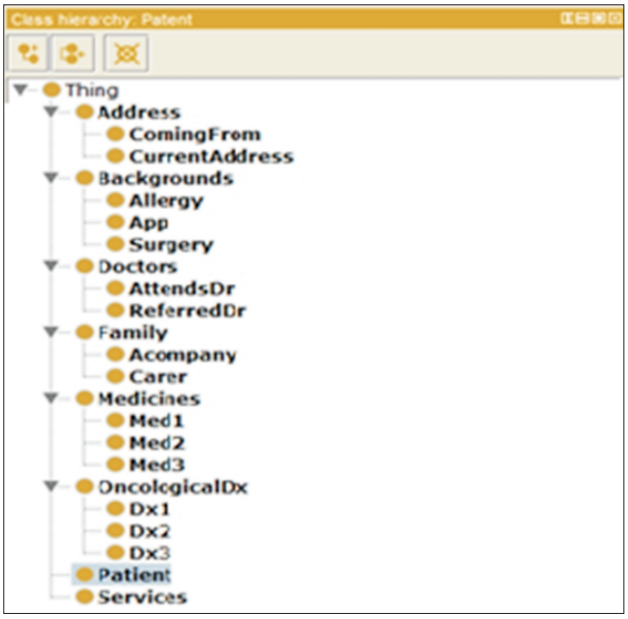

Figure 3. Excerpt from the hierarchy of classes of the domain ontology.

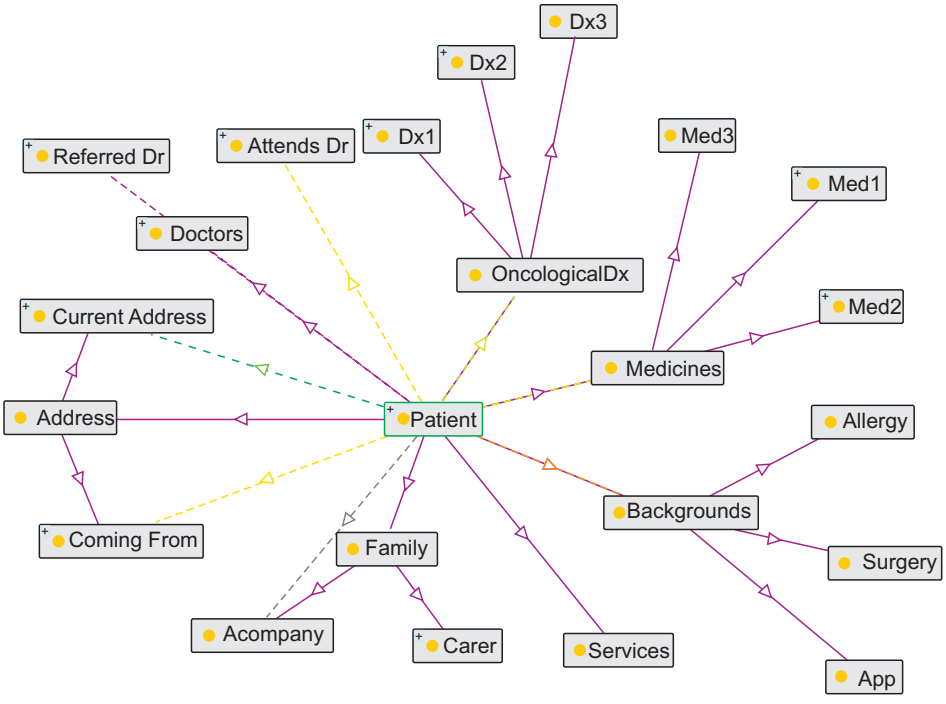

Figure 4. Excerpt from the created ontology.

https://doi.org/10.4258/hir.2018.24.4.376

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<ClinicalDocument>
<typeId extension="POCD_HD000040" root="2.16.840.1.113883.1.3"/>
<id extension="c266" root="2.16.840.1.113883.19.4"/>
<code code="11488-4" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"/>
<effectiveTime value="2015-05-06"/>
<confidentialityCode code="N" codeSystem="2.16.840.1.113883.5.25"/>
<recordTarget>
    <id extension="4-130-1598" root="2.16.840.1.113883.19.4"/>
    <name><given value="ALEJANDRO MORALES"/></name>
    <administrativeGenderCode code="MASCULINO" codeSystem="2.16.840.1.113883.5.1"/>
    <telecom value="6756-7744"/>
    <birthplace value="Chiriqui"/>
    <address value="Chepo"/>
    <service value="CIRUGIA"/>
    <doctorReferred value="Dra. Cristobalina Batista"/>
```

Figure 5. Segment of the Clinical Document Architecture file in XML format.

Table 1. Corpus metrics, experiment

Metrics	Value (%)
Precision	95.56
Recall	88.56
F-measure	91.93

tology from unstructured documents, was proposed in this study.

The developed architecture focuses on creating and populating an ontology from a text in natural language and with an appropriate level of efficiency, see Figure 4. In addition to this, the exchange of information through documents in CDA format was implemented. Figure 5 shows a segment of the CDA file in XML format.

The implemented architecture follows the functional requirements and develops the project according to the standards required by the software industry. The technological foundations presented here are those considered necessary to create a software that conveys NLP to ontologies in a clinical context.

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 so that other researches in the e-health community can modify or improve the original version of this software.

For the validation of the architecture, the project designers used the patients' clinical information domain. The total corpus consisted of the following set of data: general data (2,764 records), diagnosis (505 records), and medicines (590 records). The annotation extraction process was evaluated using the precision, recall and F-measure validation metrics shown in Table 1.

The methodology presented here has been validated in the patient clinical information domain in Spanish with promising results. It was carried out with the validation metrics most used in NLP systems, specifically the precision, recall and F-measure.

At present, there are very few ontological learning systems oriented to the information domain of patients for the construction of ontologies; therefore, research in this field is increasingly important.

To validate this methodology, a set of experiments were conducted. The system obtained satisfactory results for the annotation extraction process. The precision measures obtained for the extraction of annotations, in the corpus of palliative care, and in general medicine using the quality validation methods explained above were 95.56%, 88.56% and 91.93% which indicate a very good accuracy because they all are over 90%.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

We are grateful for the support provided by the National Secretariat of Science and Technology of Panama (SENACYT), through the National Research System (SNI), to the GISES-CIDITIC Research Group and the West Panama Regional Center of the Technological University of Panama (CRPO-UTP).

References

1. Blaya JA, Fraser HS, Holt B. E-health technologies show promise in developing countries. Health Aff (Millwood) 2010;29(2):244-51.
2. Terzieva M. Project knowledge management: how organizations learn from experience. Procedia Technol 2004;16:1086-95.
3. Maedche A, Motik B, Stojanovic L, Studer R, Volz R. Ontologies for enterprise knowledge manage-

ment. IEEE Intell Syst 2003;18(2):26-33.
4. Studer R, Benjamins VR, Fensel D. Knowledge engineering: principles and methods. Data Knowl Eng 1998;25(1):161-98.
5. Legaz-Garcia Mdel C, Menarguez-Tortosa M, Fernandez-Breis JT, Chute CG, Tao C. Transformation of standardized clinical models based on OWL technologies: from CEM to OpenEHR archetypes. J Am Med Inform Assoc 2015;22(3):536-44.
6. Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform 2013;46(5):765-73.
7. Ruiz-Martinez JM, Valencia-Garcia R, Fernandez-Breis JT, Garcia-Sanchez F, Martinez-Bejar R. Ontology learning from biomedical natural language documents using UMLS. Expert Syst Appl 2011;38(10):12365-78.
8. Inniss TR, Lee JR, Light M, Grassi MA, Thomas G, Williams AB. Towards applying text mining and natural language processing for biomedical ontology acquisition. Proceedings of the 1st International Workshop on Text Mining in Bioinformatics; 2006 Nov 10; Arlington, VA. p. 7-14.
9. Manine AP, Alphonse E, Bessieres P. Information extraction as an ontology population task and its application to genic interactions. Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence; 2008 Nov 3-5; Dayton, OH. p. 74-81.
10. Quesada-Martinez M, Mikroyannidi E, Fernandez-Breis JT, Stevens R. Approaching the axiomatic enrichment of the Gene Ontology from a lexical perspective. Artif Intell Med 2015;65(1):35-48.
11. Thakker D, Osman T, Lakin P. GATE JAPE grammar tutorial [Internet]. Sheffield: The University of Sheffield; 2009 [cited at 2018 Oct 1]. Available from: https://gate.ac.uk/sale/thakker-jape-tutorial/GATE%20JAPE%20manual.pdf.
12. IJntema W, Sangers J, Hogenboom F, Frasincar F. A lexico-semantic pattern language for learning ontology instances from text. Web Semant 2012;15:37-50.
13. Wyner AZ, Schneider J, Atkinson K, Bench-Capon TJ. Semi-automated argumentative analysis of online product reviews. COMMA 2012;245:43-50.
14. Lee CH, Wang SH. (2012). An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery. Expert Syst Appl 2012;39(10):8954-67.
15. Horridge M, Knublauch H, Rector A, Stevens R, Wroe C. A practical guide to building OWL ontologies using the Protege-OWL plugin and CO-ODE tools, Edition 1.0 [Internet]. Manchester: The University of Manchester; 2014 [cited at 2018 Oct 1]. Available from: http://mowlpower.cs.man.ac.uk/protegeowltutorial/resources/ProtegeOWLTutorialP3_v1_0.pdf
16. Fernandez-Breis JT, Chiba H, Legaz-Garcia Mdel C, Uchiyama I. The orthology ontology: development and applications. J Biomed Semantics 2016;7(1):34.
17. Lupse O, Vida M, Stoicu-Tivadar L, Stoicu-Tivadar V. Using HL7 CDA and CCD standards to improve communication between healthcare information systems. Proceedings of 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics (SISY); 2011 Sep 8-10; Subotica, Serbia. p. 453-7.

 https://doi.org/10.4258/hir.2018.24.4.376