Knowledge E
Engaging minds

Conference Paper

Ontología y Procesamiento de Lenguaje Natural

Denis Cedeño-Moreno and Miguel Vargas-Lombardo

Grupo de Investigación en Salud Electrónica y Supercomputación, Universidad Tecnológica de Panamá, Panama

Abstract

At present, the convergence of several areas of knowledge has led to the design and implementation of ICT systems that support the integration of heterogeneous tools, such as artificial intelligence (AI), statistics and databases (BD), among others. Ontologies in computing are included in the world of AI and refer to formal representations of an area of knowledge or domain. The discipline that is in charge of the study and construction of tools to accelerate the process of creation of ontologies from the natural language is the ontological engineering. In this paper, we propose a knowledge management model based on the clinical histories of patients (HC) in Panama, based on information extraction (EI), natural language processing (PLN) and the development of a domain ontology.

Keywords: Knowledge, information extraction, ontology, automatic population of ontologies, natural language processing.

Resumen

En la actualidad la convergencia de diversas áreas del conocimiento han dado lugar al diseño e implementación de sistemas de TIC que soportan la integración de herramientas heterogeneas, como la inteligencia artificial (IA), estadistica y bases de datos (BD) entre otras. Las ontologías en la computación se incluyen en el mundo de la IA y se refieren a representaciones formales de un área de conocimiento o dominio. La disciplina que se encarga del estudio y construcción de herramientas para agilizar el proceso de creación de ontologías desde el lenguaje natural, es la ingeniería ontológica. En este trabajo proponemos un modelo de gestión del conocimiento basado en las historias clinicas de los pacientes (HC) en Panamá, basados en extracción de información (EI), procesamiento de lenguaje natural (PLN) y el desarrollo de una ontologia del dominio.

Corresponding Author:
Denis Cedeño-Moreno
denis.cedeno@utp.ac.pa

Received: 15 November 2017
Accepted: 5 January 2018
Published: 4 February 2018

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the ESTEC Conference Committee.

Palabras claves: conocimiento, extracción de información, ontología, población automática de ontologías, procesamiento de lenguaje natural.

1. INTRODUCCION

El primer paso para el procesamiento informático del conocimiento lingüístico es la representación formal de dicho conocimiento. Existen múltiples recursos creados para representar la información lingüística, entre ellos, los glosarios especializados, taxonomías, tesauros y ontologías (Vilches-Blázquez 2009).

El término ontología (Choukri 2014) se ha empleado desde hace muchos siglos en el campo de la filosofía y del conocimiento. Hace ya varias décadas cobró especial relevancia en el campo de la informática. En la actualidad es parte importante dentro del ámbito de la recuperación y organización de la información y la web semántica (Bilgin et al. 2014) puesto que son clasificaciones y donde predomina la idea de transformar la red no solo en un espacio de información sino también de conocimiento.

La construcción de una ontología puede realizarse de manera manual, pero esto ocasiona diversos problemas de costo y tiempo. Como una alternativa surge el aprendizaje automático de ontologías (ontology learning) a partir de documentos textuales cuyo objetivo es identificar los elementos ontológicos de manera automática o semiautomática. Es un enfoque interesante que intenta reducir el tiempo y los recursos. Para ello se hace uso de técnicas y métodos de campos como el aprendizaje automático, la recuperación de la información o el PLN.

El objetivo principal de este trabajo consiste diseñar e implementar un método computacional que permita desde un texto escrito en lenguaje natural, es decir la HC de un paciente, extraer los elementos necesarios utilizando herramientas de PLN para luego poblar una ontología de forma automática.

El resto del documento está estructurado de la siguiente manera: Sección 2 presenta un estado del arte. Sección 3 la problemática y la solución propuesta. Sección 4 los materiales y métodos empleados. Sección 5 se describen los resultados y finalmente en la sección 6 las conclusiones y trabajo futuro.

2. ANTECEDENTES

2.1. Ontología

Menciona (Hernández 2009) que existen muchas definiciones de ontología, es un concepto muy antiguo, que viene desde el imperio griego, de ellos tenemos la definición del término (*οντος*, 'del ente', y *λόγος*, 'ciencia, estudio, teoría') y hace referencia a una rama de la metafísica que estudia la naturaleza de la existencia.

En Panamá la gran mayoría de las organizaciones de atención hospitalaria y de salud mantienen muy poca información almacenada de sus pacientes en medios electrónicos, en algunos casos esta información esta recopilada en documentos de texto.

El término ontología se ha empleado desde hace muchos siglos en el campo de la filosofía y del conocimiento y hace ya varias décadas cobró especial relevancia en el campo de la informática (Bilgin, Dikmen, & Birgonul, 2014).

Una definición muy aceptada en el área de inteligencia artificial (IA) es la de Studer (Studer, Benjamins, & Fensel, 1998), quien dijo: "Una ontología es una especificación explícita de una conceptualización".

Una ontología puede construirse de forma manual, pero representa una tarea tediosa, costosa y que consume mucho tiempo.

2.2. Trabajos relacionados

Realizar una investigación que formule una metodología de representación del conocimiento, combinando técnicas para el procesamiento de documentos textuales, herramientas de PLN y la instanciación automática de una ontología será novedosa e innovadora en áreas de convergencia como la informática y la medicina.

Consideramos entonces que esta investigación a parte de proporcionar una metodología propia de un sistema de información para toma de decisiones basado en PLN y tecnologías de representación de conocimiento, es también una fuente de documentación en tiempo real para investigadores de nuestro país.

El procesamiento de grandes volúmenes de texto libre o texto no estructurado para extraer conocimiento requiere la aplicación de una serie de técnicas de análisis entre ellas el PLN. En la actualidad se han realizado algunos trabajos relacionados que utilizan algunos de los elementos expuestos en nuestra investigación.

Como, por ejemplo, la investigación de Parisa Kordjamshidi (2015), cuya idea central es desarrollar un framework para poblar ontologías utilizando técnicas de PLN y un modelo de aprendizaje de máquina.

Cabe mencionar el trabajo que presenta un modelo semiautomático para poblar ontologías, liderado por Lennart J. Nederstigt (2014), para el dominio de e-commerce utiliza una ontología predefinida y compatible con la ontología GoogRelation.

Junto a estos enfoques tenemos la investigación de Francesco Colace (2014) que usa un sistema para el aprendizaje y población de ontologías, que combina metodologías estadísticas y semánticas.

Son varios los enfoques de investigaciones que hay en donde se combinan técnicas de PLN y el uso de ontologías de dominio para la representación del conocimiento, sin embargo, cada uno lo aplica en dominios y áreas diferentes, haciendo a cada una de estas investigaciones particularmente distintas.

3. ANÁLISIS DEL PROBLEMA

Los datos organizacionales procesados por sistemas tradicionales son almacenados en sistemas de BD, se han desarrollado incontables soluciones informáticas en la búsqueda incesante de estrategias organizacionales. A esta forma de almacenamiento de datos se le conoce como estructurada (Ricardo & Brun, 2004). Sin embargo, en muchos casos, la información de una organización la encontramos en fuentes no estructuradas, es decir escritas en lenguaje humano. Por ejemplo, en registros médicos, apuntes, actas o minutas, correos electrónicos, la web y cualquier otro medio. Día a día la creación de datos no estructurados va aumentando continuamente de manera que el análisis, procesamiento y organización de tales datos textuales se está convirtiendo en algo crucial para las organizaciones.

En nuestro país, la gran mayoría de las organizaciones de atención hospitalaria y de salud mantienen muy poca información almacenada de sus pacientes en medios electrónicos como BD, en algunos casos esta información esta recopilada en documentos de texto (Cedeño & Vargas-lombardo 2015). Las HC con los datos de los pacientes están registrados en documentos de texto y cada médico sigue sus propias normas de registro. Lo que dificulta realizar una gestión del conocimiento de sus especialistas.

3.1. Solución propuesta

Para solucionar esta problemática y poder extraer de la información no estructurada contenida en la HC del paciente el conocimiento necesario para que pueda ser utilizada por la organización, ya sea para emplearla en estrategias organizacionales o toma de decisiones, ésta requiere el uso de herramientas innovadoras que apliquen principios y técnicas de análisis de información y logren sistematizar el conjunto de conocimientos. (Villena-Román et al. 2011). Es aquí y por estas situaciones que se genera la necesidad de desarrollar metodologías con técnicas y paradigmas existentes, y la integración de métodos de análisis de datos que faciliten el proceso de exploración de documentos textuales. Una estrategia que permita completamente la preparación, el tratamiento, el análisis y visualización de información apreciable de grandes volúmenes de datos textuales y así lograr que esa información se transforme en conocimiento (Polanco, 2000).

Se propone diseñar e implementar un modelo computacional que permita desde un texto escrito en lenguaje natural, es decir la HC del paciente, poblar una ontología y luego generar un archivo OWL. Luego poder realizar razonamientos sobre esa ontología de forma tal que se extraiga conocimiento.

4. MATERIALES Y MÉTODOS

4.1. GATE

El método propuesto, utiliza técnicas de EI y PLN, para lograr la población automática de una ontología del dominio mediante la extracción de instancias desde textos escritos en lenguaje natural, se utilizará la herramienta GATE (General Architecture for Text Engineering). GATE es un software libre, escrito en Java y se utiliza para procesar texto en lenguaje natural. En la figura 1, podemos apreciar la extracción de anotaciones utilizando la herramienta GATE sobre un corpus que representa la HC de un paciente.

GATE permite definir la organización de un sistema y tiene un conjunto de procesos, los podemos apreciar en la Tabla 1.

Estos procesos que se pueden reusar, extender o adaptar, de modo que ayuda a disminuir el tiempo de desarrollo. Estos procesos toman el corpus del documento y van realizando una serie de tareas tal y como se muestra en la figura 2.

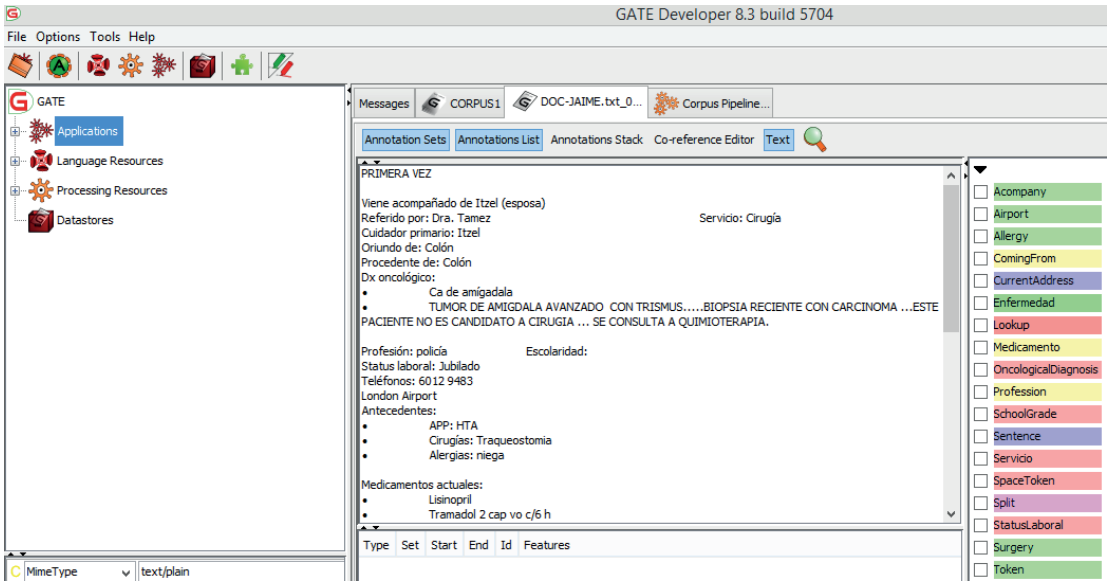

Figura 1: Recursos de GATE

Tabla 1: Explicación de los procesos del modelo

Proceso	Descripción
Corpus	Las entradas, archivos de texto, con la HC del paciente.
Document Reset PR	Consiste en reiniciar el corpus, es decir, eliminar anotaciones antiguas.
Sentence Splitter	Divide el texto en oraciones, para lo cual se utilizan transductores de estado finito.
Tokenizer	Separa las palabras que se encuentran en el texto en simples tokens.
Gazetteer	Son listas son archivos planos con una entrada por línea que se compilan mediante máquinas de estados.
NE Transducer	Se basa en técnicas de EI y también se conoce como etiquetado semántico. Recibe como parámetro un archivo "main.jape", el cual contiene una lista de gramáticas en lenguaje JAPE.

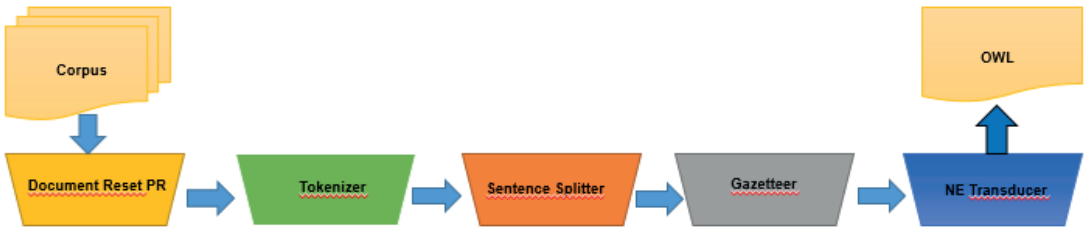

Figura 2: Procesos de GATE

4.2. JAPE (Java Annotation Patterns Engine)

GATE proporciona un lenguaje que permite reconocer entidades en un documento de texto determinado utilizando expresiones regulares llamado JAPE. GATE permite crear

los autómatas de estado finito para el total de los archivos definidos y, de esta forma, reconocer expresiones.

JAPE tiene una gramática la cual consiste en una serie de reglas patrón/acción. La parte izquierda (LHS, Left-hand-side), describe el patrón que se define para la expresión. La acción se expresa con la parte derecha de la regla (RHS, right-hand-side) y, básicamente, se utiliza para manipular anotaciones.

La regla se utiliza para encontrar la procedencia de un paciente dentro de la HC. En todas las reglas escritas en JAPE, es necesario definir un encabezado (letra color rojo).

Especificamos el nombre a la fase, para el caso de la muestra es "procedencia"; después definimos qué tipos de anotaciones se requieren como entrada, en el caso del ejemplo las anotaciones tipo "Token", "Lookup" y "Split". Luego, definimos el método que se debe aplicar en el caso de que las reglas se superpongan. Existen cinco opciones: "appelt", "first", "once", "brill", "all".

Posteriormente, se debe definir la parte izquierda (LHS) de la regla (texto color azul), en la cual se define el nombre de la regla "procedenciaRule", una prioridad para evitar ambigüedades con otras reglas y el patrón a utilizar, en nuestro caso, extraemos la procedencia del paciente, la cual está justo después de la cadena "Procedente de:".

Entonces, si se encuentra una procedencia, se asigna la etiqueta que contiene la cadena de búsqueda. Teniendo en cuenta que se cumplió el patrón, se activa la parte derecha (RHS) de la regla (texto color negro), para el ejemplo se crea una etiqueta llamada "ProcedenteDe", la cual recibe el texto que se almacenó en la etiqueta. En la figura 3 se presenta una regla en lenguaje JAPE.

```
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"}
```

Figura 3: Aplicación de una regla JAPE en la HC.

4.3. Protégé

La sección T-Box (estructura) de la ontología se hizo con la herramienta Protégé, siguiendo las normas del lenguaje ontológico OWL. En la figura 4, se muestra un extracto de la ontología realizada.

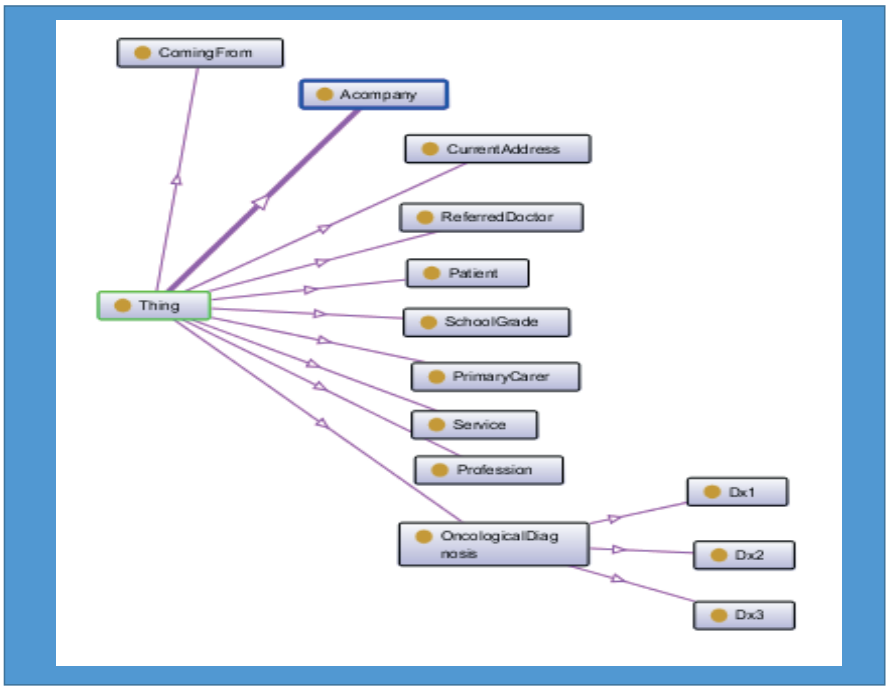

Figura 4: Extracto de la ontología en Protégé

5. DISCUSIÓN

Mediante la utilización de GATE se ha logrado etiquetar los elementos principales que componen la HC del paciente. A través de esta herramienta que nos ofrece técnicas y métodos del aprendizaje automático, logramos la recuperación de la información, programando un conjunto de reglas JAPE que van extrayendo y etiquetando la información del paciente y luego realiza las anotaciones definidas en las reglas.

Los aportes ayudan al avance de la construcción de herramientas automáticas de PLN, que puedan llevar a cabo los procesos de construcción y población de ontologías. La herramienta GATE demuestra que tiene un buen desempeño en el PLN.

Consideramos que los resultados que hemos obtenidos pueden ser mejores, si seleccionamos un número mayor de HC de pacientes para que se enriquezcan aún más los procesos de EI.

6. CONCLUSIONES Y TRABAJOS FUTÚROS

La población de una ontología puede realizarse de manera manual, pero esto ocasiona diversos problemas de costo y tiempo. Este esquema presentado representa un intento para poder reducir los procesos de creación de ontologías. La importancia de la población de ontologías, radica en que las ontologías se deben actualizar constantemente pues, de lo contrario, sería información estática. Para desarrollar los componentes del esquema de EI se utilizaron exclusivamente tecnologías de software libres lo que redunda en el hecho de no tener que pagar por licencias para el desarrollo del proyecto. En la actualidad existen muy pocos sistemas de ontology learning orientados al dominio médico para la construcción de ontologías, por ello las investigaciones realizadas en este campo es cada día más importante. Como trabajo futuro esperamos contar con un grupo más amplio de HC de pacientes para desarrollar y pulir mejor nuestro sistema y que sea eficiente a la hora de inferir conocimiento.

7. AUTORIZACIÓN Y EXONERACIÓN DE RESPONSABILIDAD

Los autores autorizan a ESTEC a publicar el artículo en los procedimientos de la conferencia. Ni ESTEC ni los editores son responsables ni del contenido ni de las implicaciones de lo que se expresa en el artículo.

Referencias

[1] Bilgin, G., Dikmen, I. & Birgonul, M.T., (2014). Ontology Evaluation: An Example of Delay Analysis. *Procedia Engineering*, 85, pp.61–68.

[2] Cedeño, D. & Vargas-lombardo, M., (2015). Framework Based on Ontologies for Palliative Care of Patients with Breast Cancer., 37(3), pp.49–57.

[3] Chang-Su Kim, Sung-Han Kim, H.-K.J., (2015). A study on web standard-based RDF converter by applying linked data and using RDF/XML standard format for data. *International Journal of Software Engineering and its Applications*, 9(1), pp.1–12.

[4] Choukri, D., (2014). A New Distributed Expert System to Ontology Evaluation. *Procedia Computer Science*, 37, pp.48–55.

[5] Corcho, O. et al., (2005). Building legal ontologies with METHONTOLOGY and WebODE. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3369 LNAI, pp.142–157.

[6] Gruber, T.R., (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2), pp.199–220.

[7] Gruber, T.R., (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5–6), pp.907–928. Available at: http://www.sciencedirect.com/science/article/pii/S1071581985710816.

[8] He, J. & Hendler, J., (2000). SHOE?: A Prototype Language for the Semantic Web 1 Introduction. *World Wide Web Internet And Web Information Systems*, pp.1–34.

[9] Hernández, A., (2009). Las ontologías: Nuevos retos. *...Perspectivas Para La ...*, pp.355–379. Available at: http://dialnet.unirioja.es/descarga/articulo/2924584.pdf.

[10] Martínez-Costa, C. et al., (2009). A model-driven approach for representing clinical archetypes for Semantic Web environments. *Journal of biomedical informatics*, 42(1), pp.150–164. Available at: http://dx.doi.org/10.1016/j.jbi.2008.05.005.

[11] McGuinness, D.L. et al., (2000). An Environment for Merging and Testing Large Ontologies. *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning KR2000*, pp.483–493. Available at: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.1812&rep=rep1&type=pdf`.

[12] Rui, L. & Maode, D., (2012). A Research on E - learning Resources Construction Based on Semantic Web. *Physics Procedia*, 25, pp.1715–1719. Available at: http://dx.doi.org/10.1016/j.phpro.2012.03.300.

[13] Ruiz-Martínez, J.M. et al., (2011). Ontology learning from biomedical natural language documents using UMLS. *Expert Systems with Applications*, 38(10), pp.12365–12378.

[14] Studer, R., Benjamins, R. & Fensel, D., (1998). Knowledge Engineering: Principles and Methods. *Data and Knowledge Engineering*, 25(1–2), pp.161–197.

[15] Vilches-Blázquez, B., (2009). Construcción de ontologías a partir de tesauros. *Semántica Espacial y descubrimiento de conocimientos para desarrollo sostenible*, pp.59–78. Available at: `http://oa.upm.es/5129/%5Cnhttp://oa.upm.es/5129/2/Construccion_de_ontologias_a_partir_de_tesauros_LMVilchesBlazquez.pdf`.

[16] Villena-Román, J. et al., (2011). Método híbrido para categorización de texto basado en aprendizaje y reglas. *Procesamiento de Lenguaje Natural*, 46, pp.35–42.