Extracción de conocimiento en documentos textuales

Extraction of knowledge in textual documents

Denis Cedeño-Moreno
Universidad Tecnológica de Panamá
Grupo de Investigación en Salud Electrónica y Supercomputación
Apartado 0819-07289 El Dorado, Panamá Provincia de Panamá, República de Panamá
denis.cedeno@utp.ac.pa

Resumen: En Panamá, existe mucha información de pacientes almacenada de forma textual, la cual no se puede manipular para gestionar un conocimiento adecuado por parte de los especialistas. Existen múltiples recursos creados para representar el conocimiento, entre ellos, los glosarios especializados, taxonomías, tesauros y ontologías. La construcción de una ontología puede realizarse de manera manual, pero esto ocasiona diversos problemas de coste y tiempo. Para resolver estos inconvenientes, se propone en esta tesis doctoral un análisis de herramientas adecuadas de procesamiento de lenguaje natural y tecnologías de representación de conocimiento para la gestión del conocimiento en documentos clínicos. El resultado de esta tesis será un sistema informático que permita instanciar una ontología del dominio que representa a los pacientes y sus enfermedades. Este enfoque se validará con un corpus de dominio médico y los resultados del proceso se medirán por los indicadores de precisión, exhaustividad y medida-F.
Palabras clave: Conocimiento, ontología, procesamiento de lenguaje natural, extracción de información

Abstract: In Panama, there is a lot of patient information stored in a textual form, which can not be manipulated to manage an adequate knowledge by specialists. There are multiple resources created to represent knowledge, including specialized glossaries, taxonomies, thesauri and ontologies. The construction of an ontology can be done manually, but this causes various problems of cost and time. To solve these problems, it is proposed in this doctoral thesis an analysis of appropriate natural language processing tools and knowledge representation technologies for knowledge management in clinical documents. The result of this thesis will be a computer system that allows to instantiate an ontology of the domain that represents the patients and their illnesses. This approach will be validated with a corpus of medical domain and the results of the process will be measured by the indicators of precision, completeness and F-measure.
Keywords: Knowledge, ontology, natural language processing, information extraction

1. Introducción

Las organizaciones modernas realizan sus actividades en un mundo globalizado, disponer del conocimiento en el momento adecuado puede suponer una clara ventaja competitiva para estar o no posicionado como la organización líder.

Desde el punto de vista de la Inteligencia Artificial (IA), el conocimiento se puede interpretar como la combinación de esquemas o estructuras de datos y procedimientos interpretativos que confieren algún comportamiento inteligente (Haugeland, 1988). Está formado por hechos, conceptos, procedimientos, ideas abstracciones, reglas y asociaciones utilizadas para modelar el mundo real.

La convergencia de diversas áreas de conocimiento actualmente ha dado lugar al diseño e implementación de sistemas informáticos que soporten la integración de herramientas innovadoras.

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.

En la actualidad, las ontologías son parte importante dentro del ámbito de la recuperación y organización de la información y la web semántica. Además, cada vez están tomando una mayor importancia dentro del PLN (Cimiano, Unger, y McCrae, 2014).

Las ontologías se pueden crear de manera manual; sin embargo, esto origina diversos problemas de costo y tiempo (Ruiz-Martínez et al., 2008).

Como una alternativa surge el aprendizaje automático de ontologías 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 IA, el aprendizaje automático (AA), la recuperación de la información (RI) o el procesamiento de lenguaje natural (PLN).

El objetivo principal de este trabajo consiste en diseñar e implementar un sistema computacional que permita desde un texto clínico escrito en lenguaje natural (LN), extraer los elementos necesarios utilizando herramientas de PLN para luego instanciar una ontología de forma automática y extraer conocimiento. Luego este conocimiento se podrá visualizar de manera amigable por parte de los sanitarios mediante un mapa conceptual.

2. Justificación de la investigación propuesta

El conocimiento, se ha convertido en el tesoro más valioso de la raza humana. Dentro de las organizaciones, pueden ser diferentes las fuentes donde encontrar este conocimiento (Terzieva, 2014).

Gran parte de este conocimiento existe en las mentes humanas y en forma de LN en libros, periódicos, informes técnicos, historias clínicas, encuestas, cuestionarios. Poder disponer de todo este conocimiento depende de nuestra habilidad para hacer ciertos procesos con la información.

Las tecnologías basadas en el conocimiento proporcionan una base coherente y fiable en las organizaciones. La gestión y visualización de ese conocimiento, desempeña un papel crucial como base tecnológica para el desarrollo de un gran número de sistemas de información (Valencia-García y Alor-Hernández, 2016).

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.

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.

3. Trabajos relacionados

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, y Birgonul, 2014).

Una definición muy aceptada en el área de IA es la de Studer (Studer, Benjamins, y Fensel, 1998 p.25), quien dijo: "Una ontología es una especificación formal y explícita de una conceptualización compartida".

Las ontologías son tecnologías que permiten una representación formal y estructurada del conocimiento donde los conceptos, las relaciones y las restricciones conceptuales son definidos mediante formalismos en un determinado dominio.

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

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.

Por su parte Suzane Santos y Rosario Girardi (2014) presentan el proyecto AppontoPro, en el método proponen un proceso incremental, para lograr la construcción y posterior población de una ontología de aplicación en el dominio de Derecho Familiar. El sistema es capaz de generar todos los elementos de la ontología tales como clases, taxonomía, relaciones no taxonómicas, instancias, propiedades y axiomas en un archivo de extensión OWL (Web Ontology Language).

Son varias las investigaciones existentes en donde se combinan técnicas de PLN y el uso de ontologías de dominio para la representación del conocimiento. Además tienen áreas de aplicación distintas, desde e-commerce, turismo, biología y otras, lo que hace que existan muchas áreas de interés sobre las cuales se puedan desarrollar nuevas investigaciones.

4. Descripción de la investigación propuesta

Las ontologías se han convertido en una importante herramienta para desarrollar aplicaciones semánticamente ricas. Los modelos ontológicos son capaces de representar una gran cantidad de información usando un pequeño número de axiomas.

Como la mayoría del conocimiento del mundo está codificado en LN, la automatización del proceso de población de las ontologías utilizando los resultados obtenidos del análisis de PLN de documentos se ha convertido recientemente en un gran desafío para aplicaciones (Witte, Khamis, y Rilling, 2010).

En Panamá casi no se gestiona la información de pacientes de forma electrónica. Esto es debido a muchos factores como la falta de presupuesto de las instituciones de salud o la falta de tiempo de los especialistas.

No existe en Panamá un método de representación del conocimiento que combine técnicas de procesos de texto o PLN gestionada por el desarrollo de una ontología y su instanciación de forma automática, para ayudar a la toma de decisiones de los especialistas u otras actividades como la investigación.

En este trabajo se propone diseñar, implementar y desarrollar un nuevo enfoque para la extracción de conocimiento a partir de la información clínica en texto en LN, basado en la utilización de herramientas de PLN para extraer información y gestionar una ontología de forma automática.

5. Metodología propuesta

Modelo propuesto: La arquitectura propuesta debe permitir extraer la información de un texto clínico escrito en LN, que representa el corpus de las historias clínicas de pacientes, extraer las entidades nombradas y elementos de conocimiento pertinentes, y generar e instanciar una ontología del dominio.

Esta ontología que contendrá la información extraída se utilizará para poder visualizar la información de una manera más conceptual y amigable para los profesionales sanitarios.

En la arquitectura propuesta las ontologías juegan un papel fundamental, ya que de su correcto diseño, estructura y complejidad dependen directamente los resultados obtenidos en los procesos.

Además, la arquitectura deberá cumplir con restricciones de interoperabilidad semántica y más concretamente deberá poder extraerse toda la información de los pacientes en el estándar de historias clínicas electrónicas HL7 (Vida, Lupse, y Stoicu-Tivadar, 2012).

Esta arquitectura se compone de varias fases que se explican a continuación:

PLN y Procesamiento del corpus: Esta fase tiene como objetivo el análisis del texto de forma lingüística. Divide el texto en oraciones y palabras.

La tarea estándar de la segmentación de las palabras se realizará con la interfaz de programación de aplicaciones proporcionada en el marco de desarrollo para el PLN llamada GATE (General Architecture for Text Engineering) (Thakker, Osman, y Lakin, 2009).

El marco de desarrollo GATE proporciona los componentes necesarios para realizar la segmentación del texto en oraciones y palabras. Estos componentes son fácilmente ensamblados para lograr una aplicación más compleja basada en tuberías, donde se agrega el componente de extracción de anotaciones. A continuación se describen estos componentes:

- Tokenizer: Realiza el proceso de separar las palabras que se encuentran en el texto en simples tokens (Berry y Castellanos, 2008). Los tokens pueden ser palabras, números, símbolos, signos de puntuación y espacio en blanco o saltos de línea.

- Sentence Splitter: Divide el texto en oraciones, para lo cual se utilizan transductores de estado finito, es decir, alfabetos de entrada y salida (IJntema et al., 2012).

Además, en esta fase, se extraen las anotaciones que luego servirán para instanciar la ontología de dominio de forma automática. Para ello, GATE proporciona una API en Java llamada StandAloneAnnie.

Para la extracción de información de las anotaciones, se utilizarán dos componentes de GATE llamado JAPE Transducer (Wyner et al., 2012) y los Gazzetteer, los cuales se encargarán de compilar y ejecutar un conjunto de reglas basadas en la gramática JAPE (Java Annotation Pattern Engine).

Instanciar la ontología: En esta fase se insertarán las instancias que poblarán nuestra ontología, en nuestra metodología las anotaciones recuperadas se utilizarán para realizar el proceso de instanciación.

La ontología de dominio será construida con Protégé (Horridge, Tsarkov, y Redmond, 2006), el cual es un editor de ontologías y de sistemas basados en conocimiento, es gratuito y de código abierto. Con esta herramienta crearemos todos los elementos de la ontología de dominio.

La ontología tendrá formato OWL, conformada por clases, subclases, propiedades e individuos. A una ontolgía, también se le conoce como base de conocimiento (knowledge base) (Guerrero et al., 2014).

OWL es un lenguaje de etiquetado semántico para publicar y compartir ontologías en la Web (Martínez-Costa et al., 2009). Se trata de una recomendación del World Wide Web Consortium (W3C) (Maldonado et al., 2012), y puede usarse para representar ontologías de forma explícita, es decir, permite definir el significado de términos en vocabularios y las relaciones entre aquellos términos u ontologías.

Las anotaciones se insertarán en la ontología como individuos de una o más clases. Para este módulo utilizaremos una API de Java llamada JENA (Zhou et al., 2010). JENA es un marco de trabajo que permite construir aplicaciones para la web semántica.

JENA tiene una serie de librerías para que los desarrolladores puedan escribir código que se encargue de procesar RDF (Resource Description Framework), OWL, SPARQL. Además incluye un motor de inferencia que se basa en reglas para razonar sobre ontologías RDF y OWL especialmente. Posee una serie de aplicaciones de almacenamiento para guardar tripletas RDF en el disco o en la memoria (Ibrahim, Mokhtar, y Harb, 2013).

Visualización de información clínica: La arquitectura que planteamos a través del sistema informático tiene como objetivo final varios escenarios en virtud a los resultados deseados, por un lado, el archivo .owl, que puede accederse desde cualquier editor de ontologías, por otro lado un archivo en formato .xml siguiendo el estándar de HL7, y tercero un reporte o interfaz gráfica que contendrá ordenadamente la información del paciente, de manera que quien la consulte pueda gestionar el conocimiento y que sirva de apoyo para la toma de decisiones.

6. Referencias bibliográficas

Bibliografía

Berry, M. W. y M. Castellanos. 2008. *Survey of text mining II*, volumen 6. Springer.

Bilgin, G., I. Dikmen, y M. T. Birgonul. 2014. Ontology evaluation: An example of delay analysis. *Procedia Engineering*, 85:61–68.

Cimiano, P., C. Unger, y J. P. McCrae. 2014. *Ontology-Based Interpretation of Natural Language*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Guerrero, J. I., C. León, I. Monedero, F. Biscarri, y J. Biscarri. 2014. Improving knowledge-based systems with statistical

techniques, text mining, and neural networks for non-technical loss detection. *Knowledge-Based Systems*, 71:376–388.

Haugeland, J. 1988. *La inteligencia artificial*. Siglo XXI.

Horridge, M., D. Tsarkov, y T. Redmond. 2006. Supporting early adoption of owl 1.1 with protege-owl and fact++. En *OWLED*.

Ibrahim, N. Y., S. A. Mokhtar, y H. M. Harb. 2013. Towards an ontology based integrated framework for semantic web. *arXiv preprint arXiv:1305.7058*.

IJntema, W., J. Sangers, F. Hogenboom, y F. Frasincar. 2012. A lexico-semantic pattern language for learning ontology instances from text. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15:37–50.

Maldonado, J. A., C. M. Costa, D. Moner, M. Menárguez-Tortosa, D. Boscá, J. A. M. Giménez, J. T. Fernández-Breis, y M. Robles. 2012. Using the researchehr platform to facilitate the practical application of the ehr standards. *Journal of biomedical informatics*, 45(4):746–762.

Martínez-Costa, C., M. Menárguez-Tortosa, J. T. Fernández-Breis, y J. A. Maldonado. 2009. A model-driven approach for representing clinical archetypes for semantic web environments. *Journal of biomedical informatics*, 42(1):150–164.

Ruiz-Martínez, J. M., J. A. Miñarro-Giménez, L. Guillén-Cárceles, D. Castellanos-Nieves, R. Valencia-García, F. García-Sánchez, J. T. Fernández-Breis, y R. Martínez-Béjar. 2008. Populating ontologies in the etourism domain. En *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volumen 3, páginas 316–319. IEEE.

Studer, R., V. R. Benjamins, y D. Fensel. 1998 p.25. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197.

Terzieva, M. 2014. Project knowledge management: how organizations learn from experience. *Procedia Technology*, 16:1086–1095.

Thakker, D., T. Osman, y P. Lakin. 2009. Gate jape grammar tutorial. *Nottingham Trent University, UK, Phil Lakin, UK, Version*, 1.

Valencia-García, R. y G. Alor-Hernández. 2016. Special issue on knowledge-based software engineering.

Vida, M., O. Lupse, y L. Stoicu-Tivadar. 2012. Improving the interoperability of healthcare information systems through hl7 cda and ccd standards. En *Applied Computational Intelligence and Informatics (SACI), 2012 7th IEEE International Symposium on*, páginas 157–161. IEEE.

Witte, R., N. Khamis, y J. Rilling. 2010. Flexible ontology population from text: The owlexporter. En *LREC*, volumen 2010, páginas 3845–3850.

Wyner, A. Z., J. Schneider, K. Atkinson, y T. J. Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. *COMMA*, 245:43–50.

Zhou, S., H. Ling, M. Han, y H. Zhang. 2010. Ontology generator from relational database based on jena. *Computer and Information Science*, 3(2):263.