Quid N° 24, pp. 69-78, enero - junio 2015, ISSN: 1692-343X, Medellín-Colombia

PROCESAMIENTO DE LENGUAJE NATURAL PARA ADQUISICIÓN DE CONOCIMIENTO: APROXIMACIONES DESDE LA INGENIERÍA DE REQUISITOS

NATURAL LANGUAGE PROCESSING FOR KNOWLEDGE ACQUISITION: APPROACHES FROM REQUIREMENTS ENGINEERING

(Recibido el 19-01-2015 - Aprobado el 20-02-2015)

PhD. Bell Manrique-Losada
Universidad de Medellín, Facultad de Ingeniería, *Docente investigadora, Grupo de Investigación* ARKADIUS*, Medellín-Colombia*
bmanrique@udem.edu.co

Resumen. En la Ingeniería de Requisitos, como fase sustancial del proceso de desarrollo de software, están involucrados procesos complejos derivados de la acción de comunicación subyacente a: la relación analista-interesado; el conocimiento que se identifica desde los stakeholders, se transforma en especificaciones intermedias y finalmente se representa formalmente; y la alta intervención de los interesados, convirtiéndolo en un proceso subjetivo. Esta complejidad en los procesos evidencia una brecha entre los discursos de interesados y analistas, exigiendo máxima intervención en la educción, en la acción de transformación de la información, y en el análisis y representación del conocimiento obtenido. En este artículo se presenta una síntesis de diversas aproximaciones reportadas en la literatura para reducir esta brecha, la mayoría de las cuales tienen como proceso común la adquisición de conocimiento desde el lenguaje natural, para la especificación de información del dominio del negocio y para su representación.

Palabras clave: procesamiento lenguaje natural, adquisición de conocimiento, ingeniería de requisitos, minería de textos, procesos de negocio

Abstract. In Requirements Engineering, as a fundamental stage in the software development process, complex processes are involved. Such processes resulted of the communication action derived from: the analyst-stakeholder relationship; the knowledge identified from stakeholders and transformed in intermediate specifications, and then in a formal representation; and, the high intervention of stakeholders. Such complexity in the processes show a gap between the analysts and stakeholders discourses, requiring maximum human participation in the elicitation, in the information transforming, and in the analysis/representation of obtained knowledge. In this article a synthesis of several approaches reported in the literature for reducing such a gap is presented. The most of them are oriented to the knowledge acquisition from natural language, for specifying and representing knowledge from business domains.

Keywords: natural language processing, knowledge acquisition, requirements engineering, text mining, business processes

1. INTRODUCCIÓN

En la Ingeniería de Requisitos, la educción es una de las fases iniciales del proceso de desarrollo de software, la cual implica obtener, analizar y especificar requisitos bajo la intervención de stakeholders, generando descripciones textuales o gráficas que plasman los conceptos más relevantes del dominio del interesado. En esta fase inicial ocurren dos sucesos que hacen complejo este proceso: la alta intervención de los stakeholders y el problema de comunicación usuario-analista. Por un lado, las técnicas de educción que se utilizan tradicionalmente requieren una muy alta intervención del interesado, debido a que en su mayoría se soportan en forma de diálogos y

Citar, estilo APA: Manrique-Losada, B. (2015). Procesamiento de lenguaje natural para adquisición de conocimiento: aproximaciones desde la ingeniería de requisitos. *Revista QUID,* (24), 69-78.

entrevistas (Gangopadhyay, 2001). Esta alta participación humana genera pérdidas de tiempo, secuencia y concisión, e incrementa los tiempos y costos del proceso de ingeniería de requisitos. Por otro lado, existe un problema de comunicación implícito en el proceso, producto de la naturaleza de la interacción usuario-interesado. La comunicación de la información obtenida en el proceso y de las descripciones del dominio de aplicación, producto del trabajo con el interesado, tienen los problemas propios del lenguaje natural [2]: considerable información sin una estructura definida, uso indiscriminado de sinónimos y ambigüedades, y redundancias de información, entre otros. Asimismo, es importante la transformación entre el lenguaje para especificar el conocimiento educido y el lenguaje natural del dominio del interesado (Zapata y Villa, 2008).

Este escenario evidencia una brecha entre los discursos de interesados y analistas, exigiendo máxima intervención en la educción, en la transformación entre la información capturada y especificaciones intermedias, en la conversión entre la descripción de requisitos y modelos de diseño y en la representación y comunicación de la información obtenida en este proceso (Coulin y Sahraoui, 2008).

En la literatura se encuentran diversas aproximaciones que intentan reducir esta brecha, desde técnicas y métodos de diferentes disciplinas. Estas aproximaciones tienen como proceso común la adquisición de conocimiento desde el lenguaje natural, tanto para la especificación de información del dominio del negocio, expectativas y necesidades de los interesados, como para su posterior representación.

En este sentido, en este artículo se presenta una síntesis de los acercamientos encontrados en la literatura hacia la adquisición de conocimiento desde la Ingeniería de Requisitos, a partir de documentos escritos en lenguaje natural.

Finalmente, bajo la categoría de técnicas híbridas, se presenta un esbozo de los enfoque particulares que se proponen hacia el procesamiento de lenguaje natural para la educción de requisitos, a partir de formalizaciones que permitan transformar documentos técnicos organizacionales escritos en lenguaje natural en algún tipo de especificación o representación de conocimiento.

El resto del artículo se organiza así: en la sección 2 se presenta el marco conceptual de ingeniería de requisitos, lenguaje natural y controlado, procesamiento de lenguaje, lingüística computacional y adquisición de conocimiento. En la sección 3 se presenta la síntesis de las aproximaciones para adquisición de conocimiento en la Ingeniería de Requisitos. Finalmente, en la sección 4 se presentan las conclusiones y el trabajo futuro.

2. MARCO CONCEPTUAL

2.1 Ingeniería de Requisitos

La Ingeniería de Requisitos es la primera fase en el desarrollo de software, cuyas actividades centrales se orientan a descubrir, capturar, analizar y especificar los requisitos que el futuro software necesita satisfacer para que se considere de calidad. Estas actividades se deben llevar a cabo de manera iterativa e incremental y priorizando por su nivel de criticidad la ejecución de los procesos de educción de requisitos y análisis de requisitos. Estos procesos involucran la identificación, adquisición y procesamiento de lenguaje natural, para su posterior representación y/o especificación en un lenguaje de modelado (Li et al., 2003).

Según Robertson y Robertson (2008) en el proceso general de ingeniería de requisitos, se desarrollan ciertas actividades, como: entendimiento del dominio de aplicación, captura y clasificación de requisitos, establecimiento de prioridades, resolución de conflictos y negociación de los requisitos del sistema. Estas actividades se consideran clave y críticas, principalmente por la acción de comunicación analista-interesado, que está orientada hacia comprender y recuperar la información esencial y relevante acerca del dominio, que se convertirá en la base de los requisitos y la cual se extrae de los interesados (cliente, usuario-final, experto del dominio).

2.2 Lenguaje Natural y Lenguaje Controlado

El lenguaje, que desde Grecia se considera esencial en la naturaleza humana, resulta poco confiable cuando la comunicación requiere ciertos niveles de precisión y cuando las acciones futuras dependen de los participantes en un proceso comunicativo. En dominios como el de la ingeniería de requisitos, donde son determinantes la exactitud y la precisión en el acto comunicativo, es importante definir reglas que determinen las relaciones entre ciertas expresiones formadas y los sentidos que ellas pretenden transmitir (Berry, 2003).

El lenguaje es un sistema usado de manera intencional por los humanos para comunicar y razonar, e involucra signos que determinan

significados (Castro et al., 2010). El lenguaje natural (LN) es el lenguaje normalmente usado por una comunidad para transmitir o expresar algo (por ejemplo información o emociones). De acuerdo con Berry (2003), la gran mayoría de requisitos están escritos en LN, por lo cual el analista requiere identificar los conceptos y relaciones entre conceptos usados por los stakeholders en un dominio, los cuales se convertirán en la base del lenguaje común entendido por analistas y stakeholders. Según Li et al. (2003), el LN es altamente informal y requiere una gran intervención del analista para el análisis y diseño de una solución software. Igualmente, dada la naturaleza ambigua del LN y los diversos factores que pueden afectar (geográficos, psicológicos y/o sociológicos) la interpretación de los textos escritos en LN, el analista debe involucrar demasiados elementos para lograr un nivel esperado de precisión en el análisis.

Un Lenguaje Controlado (LC), según Wojcik y Hoard (1995), es un subconjunto del lenguaje natural con sintaxis, semántica y/o terminología restringidas, que permite su traducción automatizada a una forma lógica (Kuhn, 2010). Haller y Schütz (2001), lo definen a partir de un conjunto de reglas que debe cumplir el lenguaje, así como el glosario que se debe utilizar.

2.3 Procesamiento de Lenguaje Natural y Lingüística Computacional

Cuando se habla de procesar un lenguaje, se refiere a la traducción de la versión de un texto desde una lengua a otra (Moreiro, 1992). Así, para ejecutar procesamiento de lenguaje natural (PLN) se requiere la transformación del texto en una representación semántica apta para razonar, tomar decisiones y ejecutar tareas específicas. Esta representación se logra por medio de procesos de Parsing o construcción de un árbol de análisis a partir de una gramática (Gavaldá, 2011). Si la gramática es sintáctica, por medio de dicho árbol de análisis se genera información sobre las categorías gramaticales de las palabras y la función sintáctica asociada (por ejemplo: identificación del sujeto, verbo, predicado y complementos). Mientras que, si la gramática es semántica, el árbol de análisis ya es bastante próximo a la representación lógica que permite el razonamiento y la ejecución. El PLN como disciplina busca desarrollar programas computacionales que sean capaces de ejecutar actividades relacionadas con la comprensión, análisis y producción de textos o discursos escritos en lenguaje natural, de una manera similar a como lo hace el ser humano (Gelbukh, 2010).

Por su parte, el lenguaje se estudia desde diferentes disciplinas orientadas hacia la lingüística, la cual a su vez estudia todos los hechos y fenómenos relacionados con el LN. Una de las disciplinas que estudia el lenguaje es la lingüística computacional, cuyo propósito es desarrollar una teoría computacional del lenguaje, a partir de las nociones de algoritmos y estructuras de datos de las ciencias de la computación (Araujo, 2006). Así, el PLN está fundamentado en la Lingüística Computacional, la cual es concebida desde la lingüística aplicada.

El término lingüística computacional (en inglés computational linguistics), se refiere al campo interdisciplinario entre la lingüística, fonética, ciencias de la computación, ciencias cognitivas, inteligencia artificial y lógica formal (Clegg, 2008). En otras palabras, según Cunningham (2000), la lingüística computacional se concentra en el estudio de lenguajes naturales, tal como lo hace la lingüística tradicional, pero usando equipos de cómputo como herramienta para modelar fragmentos de teorías lingüísticas con un interés particular.

2.4 Adquisición de conocimiento

La adquisición de conocimiento (AC) es el proceso de lograr conocimiento de un dominio (dominio del problema o dominio de una solución) a partir de: i) fuentes humanas como experto o grupo de expertos (Kendal & Green, 2006); ii) fuentes físicas como libros, documentos, sensores o archivos de computador (Turban et al., 2005).

La AC es un proceso complejo dado que implica tareas de identificación, representación, estructuración y transferencia de conocimiento a un sistema o máquina. De esta manera, dentro de este proceso son desarrolladas otras actividades relevantes relacionadas con el modelado y representación de conocimiento. Del modelado de conocimiento se generan estructuras que proporcionan un marco de trabajo para la AC y una descomposición de las tareas en sus partes más atómicas. La representación del conocimiento es a su vez una disciplina orientada hacia la simbolización del conocimiento humano en una forma específica, facilitando así el razonamiento automático. Algunos de los métodos de representación de conocimiento más comúnmente usados, son: redes semánticas, lógica, representación de conocimiento múltiple, representación bajo incertidumbre, ente otros (Al–Khanak, 2012).

La representación de conocimiento requiere que todos los elementos de información sean identificados desde el análisis y clasificados en

niveles, de acuerdo a la función que deberían cumplir en un dominio específico. Según Andrade et al. (2003) dichos niveles son: información estática, que comprende la información estructural o declarativa acerca del problema y puede ser usado en operaciones—por ejemplo conceptos, propiedades, relaciones, entre otros—e información dinámica, necesaria para conformar el comportamiento que sucede en el dominio—funcionalidad, acciones, etc.

3. APROXIMACIONES PARA ADQUISICIÓN DE CONOCIMIENTO EN LA INGENIERÍA DE REQUISITOS

En la Ingeniería de Requisitos, los procesos de adquisición de conocimiento se han desarrollado siguiendo diferentes tendencias y aproximaciones metodológicas, las cuales se han sintetizado en las siguientes categorías de análisis.

3.1 Desde el PLN

Tradicionalmente, el trabajo en PLN está centrado en ver el proceso de análisis del lenguaje en el cual están descritos los documentos fuente. Este proceso está siendo descompuesto metodológicamente en un conjunto de etapas a partir de las distinciones lingüísticas teóricas que existen entre Sintaxis, Semántica y Pragmática.

Diversos autores han definido métodos que se convierten en la base procedimental y estructural para el análisis del lenguaje de entrada. Dale (2010) define una propuesta muy cercana al proceso estándar, que se describe en términos de los siguientes pasos para ejecutarlo:

El primer paso es la etapa del preprocesamiento del texto, que incluye tokenización y segmentación de oraciones. Los textos en lenguaje natural (documentos técnicos) generalmente no son cortos, ordenados o bien organizados, por lo cual se requiere la separación de sus oraciones como unidades de análisis de los textos.

La segunda etapa es el procesamiento del texto. Este incluye el análisis léxico, sintáctico y semántico. El análisis léxico es donde ocurre la descomposición a grano fino de las palabras, a partir de la identificación de cada una de sus partes siguiendo reglas léxicas predefinidas. El análisis sintáctico por su parte implica la extracción de sentido de las oraciones, analizando cada oración y determinando su estructura. Finalmente, el análisis semántico permite identificar el significado de cada oración bajo análisis.

La última etapa está relacionada comúnmente con el procesamiento contextual, el cual busca analizar elementos del contexto en el que se desarrollan o se usan los textos.

A su vez, dentro de esta categoría de análisis, algunas de las propuestas y tendencias de adquisición de conocimiento más representativas desde el PLN para la Ingeniería de Requisitos, son:

Análisis de información contenida en documentos sin modificación:

O'Shea y Exton (2004) propone una técnica de extracción de requisitos desde un corpus de reportes de errores.

Meth et al. (2012) propone un método de apoyo al proceso de educción de manera automatizada.

Análisis de documentos semiestructurados:

Cybulski y Reed (1998), en el marco del proyecto RARE, analiza textos de requisitos por medio de una red semántica asistida por un tesauro.

Cleland-Huang et al. (2008) propone un método de detección de puntos de vista desde documentos de requisitos no funcionales.

Bajwa et al. (2011) realiza el análisis de textos de reglas de negocio, para su conversión a un vocabulario semántico de negocios.

Antón (1997) identifica requisitos no funcionales desde funcionales, aplicando técnicas de procesamiento.

Hahn et al. (1996) integra técnicas de PLN en un método para identificar conceptos y conocimiento desde textos escritos en alemán (reportes de pruebas sobre productos tecnológicos y reportes médicos).

Young y Antón (2010) proponen la identificación de requisitos desde documentos normativos y de políticas, analizando compromisos, privilegios y derechos.

Clements y Northrop (2002) proponen técnicas para revisar documentos del proceso de líneas de productos software para identificar características de productos.

Kray y Porzel (2000) definen un método de diseño de agentes para el análisis semántico y pragmático profundo de conceptos espaciales de textos de esta disciplina, los cuales son expresados en una especificación semiformal.

Lee y Bryant (2002) presentan un lenguaje de marcado orientado por agentes (DAML en el Proyecto DARPA), el cual permite la generación de expresiones de salida como representaciones formales de requisitos escritos de manera informal.

3.2 Desde la Ingeniería del Conocimiento

La ingeniería del conocimiento se orienta hacia la definición de un conjunto de reglas que codifican el conocimiento para analizar información siguiendo ciertas categorías dadas.

En la literatura se encuentran varios aportes desde la ingeniería del conocimiento hacia el dominio de la ingeniería de requisitos o procesos relacionados, como se presenta a continuación.

3.2.1 Técnicas de la Ingeniería Ontológica

La ingeniería ontológica es una disciplina orientada hacia la creación de sistemas que permitan compartir, reutilizar y analizar conocimiento sobre un dominio dado, a partir, según Gómez et al., (2004), de representaciones jerárquicas de conceptos y sus clasificaciones asociadas. Además, se convierten en herramientas de referencia para la construcción de sistemas de bases de conocimiento que aporten consistencia, fiabilidad y falta de ambigüedad a la hora de recuperar información.

Siguiendo esta orientación, cualquier propuesta que busca generar dichas representaciones a partir de conceptos de un dominio, está aplicando técnicas o métodos de la ingeniería ontológica. En el domino de la Ingeniería de Requisitos, se han realizado aportes como los siguientes:

Aussenac-Gilles et al. (2000) proponen la validación de documentos legislativos a partir de análisis ontológico. Con esta misma orientación, autores como Dinesh et al. (2006) han definido métodos para validar el cumplimiento de ciertos documentos organizacionales bajo ciertas regulaciones. Rösner et al. (1997) usan técnicas para analizar bases de conocimiento y generar documentos multilingües en dominios específicos. Algunas de estas técnicas son basadas en lógica formal, lo cual permite realizar inferencias lógicas mediante la utilización de una serie de componentes como la inclusión de axiomas, lógica de primer y segundo orden, etc.

Algunas propuestas con un mayor acercamiento a la lógica formal, son: parametrización semántica, como proceso para representar descripciones del dominio en lógica de predicados de primer orden (Breaux et al., 2006). Vegega et al. (2012) define métodos de formalización de dominios de negocio para grandes proyectos de explotación de información, basado en técnicas de ingeniería del conocimiento.

3.2.2 Técnicas de Minería de datos y de textos

La minería de textos o text mining (por sus siglas en inglés) se refiere a la exploración de colecciones de documentos y el descubrimiento de información no contenida en ningún documento individual, pero sí producto de la integración de varios (Nasukawa y Nagano, 2001). Las técnicas de minería de textos se sustentan en el hecho que la gran mayoría de información de las organizaciones está almacenada en forma de documentos. Sin duda, este campo de estudio es muy vasto, pues es apoyada por técnicas de áreas relacionadas, como la categorización de textos, el procesamiento de lenguaje natural, la recuperación de la información y el aprendizaje de máquina.

Técnicas de minería de textos permiten extraer conocimiento desde grandes cantidades de datos desde textos. Este tipo de técnicas, de manera similar a como se hace con la recuperación de información, permite seleccionar documentos que cumplan con las expectativas o intereses de un usuario, así como también encontrar patrones y reglas factibles en textos que indican tendencias y características significativas acerca de aspectos o temas específicos. Mooney y Bunescu (2005) identifican conocimiento abstracto desde un corpus de textos y conocimiento concreto desde un conjunto de datos, utilizando técnicas de minería de datos tradicionales que descubren patrones de uso generales. Esta identificación la realizan sobre textos de resúmenes biomédicos reales, anuncios y descripciones de productos.

3.3 Aproximaciones metodológicas desde el aprendizaje de máquina

Aysolmaz y Demirors (2014) introducen un método para realizar especificación de requisitos desde modelos de procesos de negocio, y su representación en forma de diagramas o documentos de requisitos. En este método se especifican procedimientos que contienen metadatos y descripciones de los modelos y se generan de manera automática documentos de procesos.

Da Cunha et al. (2014) proponen una descripción formal, en términos de un metamodelo, para representar el contexto de las actividades de un proceso de negocio. Este metamodelo define el lenguaje y los procesos en términos de una colección de conceptos del dominio. Esta propuesta integra

técnicas de análisis de elementos contextuales con formalismos basados en lógica, para implementar la generación de una representación formal de un contexto.

Byeong-Ho y Richards (2010) muestran resultados de exploración de documentos y proponen un explorador de casos que permite refinar y mejorar de manera manual un conjunto de reglas para identificación de conocimiento desde textos. A partir de técnicas de aprendizaje de máquina se logra la inferencia de nuevas reglas. Algunos otros autores como Riedmiller y Merke (2002) y Akiyama (2007) han logrado la adaptación de técnicas de aprendizaje de máquina como aprendizaje por refuerzo al dominio de adquisición de conocimiento. Otros investigadores han logrado la combinación de técnicas para reducir la cantidad de intervención de expertos en procesos de adquisición de conocimiento (Suryanto y Compton, 2003).

3.4 Técnicas híbridas apoyadas en recursos de análisis del lenguaje

Incluye aquellas técnicas y recursos para la creación, adaptación y modelado del lenguaje subjetivo e informal generado en diversas fuentes de información (blogs, redes sociales, documentos técnicos, etc.). El uso de estos recursos de análisis del lenguaje es un medio adecuado para describir y analizar textos en un área específica.

3.4.1 Basadas en Corpus

El uso de corpus como recurso de análisis de lenguaje es un medio muy adecuado para describir y analizar textos en un área dada. Las ventajas son muchas en comparación con la intuición y conocimiento de expertos del dominio. Según Krishnamurthy (1997), algunas de dichas ventajas son que el corpus puede ser más entendible y balanceado y poder generar estadísticas de manera más precisa y rápida.

Autores como Calzolari (1997) y Aston (1997) se apoyan en procesamientos computacionales de corpus para analizar documentos organizacionales.

Wang (2005) propone la integración de métodos de teoría de conjuntos con procesamiento basado en corpus, buscando crear bases de conocimiento sin la intervención de un ingeniero de conocimiento o un experto. A partir de esta propuesta se han desarrollado nuevos métodos para interpretar significados y determinar otras formas de conocimiento basados en heurísticas/reglas, como razonamiento basado en casos y razonamiento basado en marcos.

3.4.2 Basadas en técnicas de PLN e ingeniería del conocimiento

Zolotarev et al. (2012) desarrollan un método para construir una estructura formalizada de un dominio, basado en análisis de textos en lenguaje natural. Este método incluye descubrimiento de objetos, identificación de propiedades y acciones relacionadas, descubrimiento de procesos de negocio de un dominio específico, formación de un tesauro específico del dominio y finalmente, la especificación de nuevos procesos de negocio.

Hahn y Schnattinger (1998) proponen una metodología para adquisición de conocimiento basado en texto en la cual bases de conocimiento de dominio continuamente son mejoradas como producto del proceso de comprensión de un texto. Los conceptos son adquiridos tomando en cuenta dos fuentes de evidencia: conocimiento del domino de los textos que sirve como escala de comparación para juzgar la posibilidad de nuevas descripciones de conceptos derivados a la luz del conocimiento previo; y los patrones lingüísticos que permiten evaluar la intensidad de la fuerza interpretativa que puede ser asignada a las construcciones gramaticales en las cuales ocurren los elementos léxicos.

Charnine y Somin (2014) presentan un método de búsqueda semántica de textos web escritos en LN para la extracción de información temática y conceptos clave. Esta información de salida es procesada por un módulo de análisis estadístico para la formación de un perfil de dominio temático que contiene asociaciones entre términos del dominio, palabras clave y frases relevantes.

3.4.3 Basadas en minería de textos

De manera similar a las técnicas basadas en corpus, las técnicas de minería de textos están siendo usadas de manera incremental para extraer información probablemente útil desde grandes cantidades de textos y analizando sus patrones de uso.

Nasukawa y Nagano (2001) han desarrollado sistemas que sirven de referencia en este proceso, que intentan analizar textos y explorar conocimiento, para su posterior representación. Estos autores hacen una mezcla de técnicas de minería de textos con técnicas de análisis estadístico de los resultados generados luego de los procesos de análisis. Murakami y Nasukawa (2004) proponen un método para hallar expresiones sinónimas a partir de un conjunto de corpus e hipótesis de uso y significado de expresiones. Este método es útil para caracterizar contextos y crear bases terminológicas en las

organizaciones. Fonseca et al., (2007) presentan una propuesta de minería de textos, soportada en gestión de corpus, para extraer conocimiento usando un modelo de clasificación desde datos no estructurados de un sistema diagnóstico. De manera particular, esta propuesta integra además métodos de aprendizaje de máquina para crear reglas de inferencia a partir de los hallazgos que se vayan logrando en el proceso de análisis.

Una propuesta de ingeniería ontológica e ingeniería de textos presentada por Quasthoff y Wolff (2002) describe cómo extraer relaciones semánticas relevantes desde colecciones de textos. Zouaq y Nkambou (2008) proponen una metodología semiautomática para construir ontologías del dominio desde textos escritos en inglés. Para generar esta ontología, se pasa por unos modelos de conocimiento intermedios (mapas de conceptos) que luego son representados en una ontología OWL, siguiendo los pasos de procesamiento típicos del PLN. Fernández et al. (2008) usan técnicas de análisis lingüístico para especificar grafos conceptuales desde textos no estructurados. Entre las técnicas usadas está la minería e interpretación de contenidos, clustering y cleaning de datos. El proceso de adquisición de conocimiento que se lleva a cabo es desarrollado sin supervisión y minimizando la cantidad de fallas en la información recuperada desde los documentos.

4. CONCLUSIONES Y TRABAJO FUTURO

Para reducir la brecha existente entre los discursos de los interesados y los analistas en el proceso de ingeniería de requisitos, diversos investigadores han logrado aproximaciones en términos de técnicas y métodos de diferentes disciplinas. Estas aproximaciones tienen como proceso común la adquisición de conocimiento desde el lenguaje natural hacia una especificación formal de información del dominio del negocio, expectativas y necesidades de los interesados. Estos aportes se pueden clasificar en las siguientes categorías: i) Procesamiento de Lenguaje Natural; ii) Ingeniería del Conocimiento; iii) Aprendizaje de Máquina; y iv) Técnicas Híbridas.

En este artículo se presenta una síntesis de los acercamientos encontrados en la literatura hacia la adquisición de conocimiento desde la Ingeniería de Requisitos, a partir de documentos escritos en lenguaje natural. Estos acercamientos catalogados en tres grandes perspectivas, finalmente son sintetizados en una categoría final de 'técnicas híbridas', que refleja los aportes más integradores y cercanos a la reducción de la brecha de comunicación que existe en el proceso de Ingeniería de Requisitos. La especificación esperada a partir de la integración de estas técnicas y perspectivas híbridas, se verá reflejada en un proceso de transformación de grandes corpus de textos, a partir de patrones funcionales, estructurales y lingüísticos de textos que serán logrados aplicando técnicas de aprendizaje de máquina, para obtener conocimiento del dominio organizacional e información del negocio, útil para el proceso de ingeniería de requisitos.

RECONOCIMIENTO

Este trabajo se enmarca dentro del proyecto de investigación "Especificación de un Lenguaje Controlado de Dominio Específico: Fundamentos Lingüísticos y bases de Transformación desde Documentos Técnicos Corporativos en Lenguaje Natural". Proyecto aprobado en "Programa Nacional de Proyectos para el fortalecimiento de la investigación, la creación y la innovación en Posgrados de la Universidad Nacional de Colombia 2013-2015" y Convocatoria No. 33 de la Universidad de Medellín. Entidades: Universidad Nacional de Colombia, Universidad de Medellín y Wake Forest University.

REFERENCIAS

Akiyama, H. (2007). Team Description RoboCup 2007. *LNCS (LNAI)*. Heidelberg: Springer.

Al–Khanak, S. A. K. (2012). University experts' knowledge acquisition: Investigating knowledge engineering system approach. *African Journal of Business Management* 7(2), 135-143.

Andrade, J., Ares, J., García, R., Pazos, J., Rodríguez S. y Silva, A. (2003). A methodological framework for generic conceptualization: problem-sensitivity in software engineering. *Information and Software Technology,* 46(10), 635–649.

Antón, A. I. (1997). *Goal Identification and Refinement in the Specification of Software-Based Information Systems.* (Tesis de Doctorado) Georgia Institute of Technology, Estados Unidos.

Araújo, L. (2006). *Procesamiento de Lenguaje Natural.* Recuperado de http://tabasco.torreingenieria.unam.mx/gch/PLN/cap1.pdf

Aston, G. (1997). Enriching the Learning Environment: Corpora in ELT. En Wichmann et al. (Eds.), *Teaching and Language Corpora*. London: Longman Limited.

Aussenac-Gilles, N. Biébow, B. y Szulman, S. (2000). Revisiting Ontology Design: A Method Based on Corpus Analysis. *Knowledge Engineering and Knowledge Management. Methods, Models, and Tools: 12th International Conference* (EKAW), Juan-les-Pins, France.

Aysolmaz, B. y Demirors, O. (2014). Modeling business processes to generate artifacts for software development: a methodology. *Proceedings of the 6th International Workshop on Modeling in Software Engineering* (MiSE). NY, Estados Unidos.

Bajwa, I. S. Lee, M. y Bordbar, B. (2011). SBVR Business Rules Generation from Natural Language Specification. *AAAI. Artificial Intelligence for Business Agility*. Spring Symposium.

Berry, D. M. (2003). Natural language and requirements engineering - Nu? [Presentación de clase] CSD & SE Program University of Waterloo, Recuperado de http://www.ifi.unizh.ch/groups/req/IWRE/papers&presentations/Berry.pdf

Breaux, T., Antón, A. y Doyle, J. (2008). Semantic parameterization: A process for modeling domain descriptions. ACM Trans. *Softw. Eng. Methodol*. 18, 27.

Byeong-Ho, K. y Richards, D. (2010) Knowledge Management and Acquisition for Smart Systems and Services. *11th International Workshop, PKAW 2010,* Computer Science Springer. Daegu, Korea.

Calzolari, N. (1997). *Lexicon and Corpus: A Multi-faceted Interaction*. Cicle de Conferènces 95-96. Lèxic, Corpus i Dictionaris. Barcelona: Institut Universitari de Lingüística Aplicada.

Castro, L., Baiao, F. y Guizzardi, G. (2009). A survey con Conceptual Modeling from a Linguistic Point of View. *Relatórios Técnicos do Departamento de Informática Aplicada da UNIRIO,* (19), 3-12.

Charnine, M., Somin, N. Nikolaev, V. (2014). Conceptual text generation based on key phrases. *Proceedings of the 2014 International Conference on Artificial Intelligence*. CSREA Press. Las Vegas, Estados Unidos.

Clegg, A. (2008). *Computational-Linguistic Approaches to Biological Text Mining*. (Tesis de Doctorado)Escuela de Cristalografía, University of London. Inglaterra.

Cleland-Huang, J., Marrero, W. y Berenbach, B. (2008). Goal-Centric Traceability: Using Virtual Plumblines to Maintain Critical Systemic Qualities. *IEEE Transactions on Software Engineering,* 34(5). 685 – 699.

Clements, P. y Northrop, L. (2002). Software Product Lines: Practices and Patterns. Boston, MA: Addison-Wesley.

Coulin, Ch. y Sahraoui, A. (2008). A Meta-Model Based Guided Approach to Collaborative Requirements Elicitation. *Proceedings of the 18th annual international symposium of INCOSE,* 1-6.

Cunningham, H. (2000). *Software Architecture for Language Engineering.* (Tesis de Doctorado) Departamento de ciencias de la computación, University of Sheffield. Reino Unido.

Cybulski, J. y Reed, K. (1998). Requirements Classification and Reuse: Crossing domains boundaries. *6th Intl. Conf. on Software Reuse (ICSR'2000)*. Springer, Vienna, Austria.

Da Cunha, T., Santoro, F. M., Revoredo, K., Tavares, V. (2014). A formal representation for context-aware business processes. *Computers in Industry,* 65(8), 1193-1214.

Dale, R. (2010). Classical Approaches to Natural Language Processing. **En** Nitin I., Damerau F. (Eds.) *Handbook of Natural Language Processing,* Estados Unidos, CRC Press.

Dinesh, N., Joshi, A., Lee, I., y Webber, B. (2006). Extracting formal specifications from natural language regulatory documents. *ICoS-5,* Buxton, England.

Fernández, M., de la Clergerie, E. y Vilares, M. (2008). Mining conceptual graphs for knowledge acquisition. *Proceedings of the 2nd ACM workshop on improving non-English web searching (iNEWS '08).* ACM, NY, USA.

Fonseca Silveira, S. M., Marchi, R., Cunha da Silva, L. M., Sampaio de Souza, K. X., Mendonça de Oliveira, L. H., de Medeiros Oliveira, S. R. y Morand, M. A. (2007). *An approach based on text mining for knowledge acquisition in diagnostic systems*. Embrapa, Brasil.

Gangopadhyay, A. (2001). Conceptual Modeling from Natural Language Functional Specifications. *Artificial Intelligence in Engineering*, 15(2), 207-218.

Gavaldá, M. (2011). *La investigación en tecnologías de la lengua. Research in language technology*. Recuperado de http://quark.prbb.org/19/019021.htm

Gelbukh, A. (Ed.) (2010). Natural Language Processing and its Applications. *Research in Computing Science*, 46, 3-15.

Gómez-Pérez, A., Fernández-López M. y Corcho, O. (2004). *Ontological Engineering*. International: Springer.

Hahn, U., Klenner, M. y Schnattinger, K. (1996). A Quality-Based Terminological Reasoning Model for Text Knowledge Acquisition. **En** O'Hara & Schreiber (Eds.). *Advances in Knowledge Acquisition. Shadbolt*, Berlin: Springer-Verlag.

Haller, J. y Schütz, J. (2001). CLAT: Controlled Language Authoring Technology. *Proceedings of the 19th Annual international Conference on Computer Documentation*, Santa Fe, Estados Unidos.

Hans, U. y Schnattinger, K. (1998). Towards Text Knowledge Engineering. *AAAI-98 Proceedings.*

Kang, B. H. y Richards, D. (2010). RDRCE: Combining Machine Learning and Knowledge Acquisition. Knowledge Management and Acquisition for Smart Systems and Services. *Lecture Notes in Computer Science* 6232, 165-179.

Kendal, S. y Creen, M. (2006). *An Introduction to Knowledge Engineering,* International: Springer.

Kray, C. y Porzel, R. (2000). Spatial cognition and natural language interfaces in mobile personal assistants. *Proceedings of the ECAI 2000 Workshop on Artificial Intelligence in Mobile Systems.* Berlin, Germany.

Krishnamurthy, R. (1997). *Keeping Good Company: Collocation, Corpus and Dictionaries.* Cicle de Conferènces 95-96. Lèxic, Corpus i Dictionaris. Barcelona: Institut Universitari de Lingüística Aplicada.

Kuhn, T. (2010). *Controlled English for Knowledge Representation*. (Tesis de Doctorado) Faculty of Economics, Business Administration and Information Technology, University of Zurich. Suiza.

Lee, B. y Bryant, B. R. (2002). Contextual Natural Language Processing and DAML for Understanding Software Requirements Specifications. *COLING 2002: The 19th International Conference on Computational Linguistics.* Taipei, Taiwan.

Li, K. Dewar R. G. y Pooley R. J. (2003). *Requirements capture in natural language problem statements*. Recuperado de http://www.macs.hw.ac.uk/cs/techreps/docs/files/HW-MACS-TR-0023.pdf

Meth, H., Li, Y. Maedche, A. y Mueller, B. (2012). Advancing task elicitation systems –an Experimental evaluation of design Principles. *Proceeding 33 International Conference on Information Systems*, Orlando Florida, Estados Unidos.

Mooney, R. J. y Bunescu, R. (2005). Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl*. 7(1), 3-10.

Moreiro, J. (1992). Perspectiva Documental del Procesamiento de Lenguaje Natural. *Memorias Congreso SEPLN VIII,* Universidad Carlos III, Madrid.

Murakami, A. y Nasukawa, T. (2004). *Term Aggregation: Mining Synonymous Expressions using Personal Stylistic Variations.* Tokyo Research Laboratory.

Nasukawa, J. y Nagano, A. (2001). Text analysis and knowledge mining system. *IBM Systems Journal*, 40(4), 967-984.

O'Shea, P. y Exton, C. (2004). The Application of Content Analysis to Programmer Mailing Lists as a Requirements Method for a Software Visualization Tool. *12th Internat. Works on Software Technology Practice*, 30–39.

Quasthoff, U. y Wolff, W. (2002). *Text-based Knowledge Acquisition for Ontology Engineering*. En Konvens 2002, Universidad de Leipzig.

Riedmiller, M., Merke, A. (2002). Using Machine Learning Techniques in Complex Multi-Agent Domains. **En** Stamatescu, I., Menzel, W., Richter, M., Ratsch, U. (eds.) *Perspectives on Adaptivity and Learning. LNCS*. Heidelberg: Springer.

Robertson, S. y Robertson, J. (2008). *Mastering the Requirements Process,* Estados Unidos: Addison Wesley.

Rosner, D., Grote, B., Hartmann, K. y Hofling, B. (1997). From Natural Language Documents to Sharable Product Knowledge: A Knowledge Engineering Approach. *Journal of Universal Computer Science,* 3(8), 955-987.

Suryanto, H., Compton, P. (2003). Invented Knowledge Predicates to Reduce Knowledge Acquisition Effort. **En** Tecuci, G., Aha, D., Boicu, M., Cox, M., Ferguson, G. (eds.) *Proceedings of the IJCAI 2003 Workshop on Mixed-Initiative Intelligent Systems, Eighteenth International Joint Conference on Artificial Intelligence,* Austin.

Turban, E., Aronson, J. y Liang, T. (2005). Knowledge Acquisition, Representation, and Reasoning. Chapter 18. **En** Decision Support Systems and Intelligent Systems. Estados Unidos: Prenhall Editors.

Vegega, C., Amatriain, H., Pytel, P., Pollo-Cattaneo, F., Britos P., y Garcia-Martinez, R. (2012). Formalización de Dominios de Negocio basada en Técnicas de Ingeniería del Conocimiento para Proyectos de Explotación de Información. *Proceedings IX Jornadas Iberoamericanas de Ingeniería del Software e Ingeniería del Conocimiento*.

Wang, F. H. (2005). On Acquiring Classification Knowledge from Noisy Data Based on Rough Set. *Expert Systems with Applications*. 29(1). 9-64.

Wojcik, R. y Hoard, J. (1995). *Controlled Languages in Industry* Recuperado de http://www.cslu.ogi.edu/HLTsurvey/ch7node8.html

Young, J. D. y Antón, A. I. (2010). A Method for Identifying Software Requirements Based on Policy Commitments. *18th IEEE International Requirements Engineering Conference*.

Zapata, C.M., y Villa, F. A. (2008). La Gramática Básica de UN-Lencep expresada en HPSG. *Revista Avances en Sistemas e Informática*, 5(1), 81-92.

Zolotarev, O., Charnine, M. y Matskevich, A. (2012). Conceptual business process structuring by extracting knowledge from natural language texts. *Proceedings of WORLDCOMP'12,* 1. New University, Russian.

Zouaq, A. y Nkambou, R. (2008). Building Domain Ontologies from Text for Educational Purposes. *Learning Technologies, IEEE Transactions* 1(1), 49-62.