Transformación de lenguaje natural a lenguaje controlado para la educción de requisitos a partir de documentación técnica

Turning natural language into controlled language in order to educe requiriments from technical documentation

Bell Manrique Losada
Estudiante de Doctorado en Ingeniería, Profesora Asistente
Universidad de Medellín, bmanrique@udem.edu.co

Carlos Mario Zapata Jaramillo
Ph.D. en Ingeniería, Profesor Asociado Universidad Nacional de Colombia, sede Medellín, cmzapata@unal.edu.co

Resumen
En la ingeniería de software, la educción de requisitos es el proceso mediante el cual un analista intenta capturar los requisitos que tiene un interesado respecto de un futuro aplicativo de software. Este proceso implica una traducción de un lenguaje natural—del interesado—a un lenguaje controlado. Tradicionalmente, para obtener estos requisitos se utilizan técnicas relacionadas con entrevistas y diálogos, lo cual genera grandes esfuerzos de los analistas y pérdida de tiempo, por la complejidad que implica el procesamiento del lenguaje natural. No es común educir requisitos a partir de fuentes indirectas como documentación técnica (manuales de procedimientos, reglamentos y estatutos, entre otros). En este artículo se contextualiza la problemática asociada con la educción de requisitos y las brechas que existen entre los lenguajes controlados, para especificar requisitos, y el lenguaje natural del interesado. Finalmente, se describe el escenario que se plantea lograr con la puesta en marcha de estas propuestas, a partir de la revisión de literatura realizada.
Palabras clave: documentación técnica, educción de requisitos, lenguaje controlado, lenguaje natural.

Abstract
Analysts try to capture the stakeholder requirements related to a future software application through a requirements elicitation, a branch of software engineering. Requirements elicitation implies a translation

Recibido: 11 de junio 2011
Aprobado: 26 de octubre 2011

from a natural language—the stakeholder's—into a controlled language. Commonly, analysts use techniques related to dialogues and interviews in order to meet such requirements. These techniques demand a huge amount of analysis and a waste of time, due to the complexity of natural language processing. Eliciting requirements from indirect sources, like technical documentation (i.e., procedure manuals, regulations, and statutes) is not a common task. In this paper we contextualize the problems linked to the requirements elicitation process and the existing gaps between the controlled language—to specify requirements—and the stakeholder's natural language. Finally, we describe a research scenario proposed, based on a state-of-the-art review.
Keywords: controlled language, natural language, requirements elicitation, technical documentation.

Introducción

La educción es una de las primeras fases de la *ingeniería de requisitos* para el desarrollo de software, cuyo propósito principal es descubrir todos los requisitos que el futuro software necesita satisfacer, para que alcance los objetivos definidos (Cheng & Atlee, 2007). Se relaciona principalmente con la acción de comunicación analista-interesado, que busca recuperar información esencial y relevante acerca del dominio (expresada en lenguaje natural). Esta información se convertirá en la base de los requisitos y se captura con los interesados (cliente, usuario-final, experto del dominio, etc.; Kof, 2004). En este proceso son determinantes la exactitud y la precisión en los discursos en lenguaje natural, lenguaje en el cual, según Berry (2003), se escribe la gran mayoría de requisitos. Para el analista es muy importante identificar los conceptos y las relaciones entre conceptos que emplea el interesado, pues ellos se convertirán en la base del lenguaje común que deben entender el analista y el interesado. Lo anterior requiere, según Li *et al.* (2003), mayor intervención y transformación para las tareas posteriores de análisis y diseño del producto software.

Este proceso se puede visualizar como una 'traducción' de un lenguaje a otro, de manera tal que el analista (traductor) debe reconocer y entender símbolos expresados en un lenguaje natural de un universo de discurso (del interesado) y transformarlos en un conjunto de símbolos definidos en un lexicón de un lenguaje controlado (Castro *et al.,* 2009). Posteriormente, el analista debe representar estos símbolos en lenguajes técnicos (generalmente de tipo gráfico, como los esquemas conceptuales). Los interesados, por su parte, validan los requisitos capturados y plasmados en dichos modelos y esquemas conceptuales, a pesar de no comprenderlos suficientemente, pues los lenguajes técnicos se alejan del natural. Esta comunicación entre los participantes (interesado-analista) se torna más débil aún por las diferencias de formación y experiencia entre ellos, lo que genera problemas en la correcta validación de los requisitos, en el alcance de los modelos conceptuales generados y, finalmente, en los costos de ejecución del proceso de educción (Bolton *et al.,* 1994).

A partir de las técnicas de captura de requisitos utilizadas de forma tradicional (Christel & Kang, 1992), la intervención del interesado se suele estimar incorrectamente, pues la mayoría de las veces se realiza en forma de diálogos y entrevistas (Leite, 1987) con la resultante pérdida de tiempo, costos, coherencia, concisión, entre otros, como se verá más adelante. A pesar de los acercamientos propuestos en la literatura para reducir la brecha entre los universos de discurso del interesado y del analista, todavía se exige máxima intervención del analista en el proceso de educción, en la conversión entre la descripción de requisitos y modelos de diseño y en el método que guíe este proceso. Es necesario, entonces, lograr acercamientos entre los lenguajes controlados existentes para especificar los requisitos y el lenguaje natural del interesado. Este artículo presenta el área problemática que describe este escenario, la revisión de literatura asociada con el problema principal identificado y los resultados que se espera lograr con la ejecución de tales acciones.

El resto del artículo se organiza así: la Sección 2 describe el área problemática y el marco teórico-conceptual que lo sustenta; la Sección 3 presenta una serie de problemas identificados y, finalmente, el

problema de investigación que se propone abordar; la Sección 4 muestra, a manera de conclusiones, los resultados que se espera lograr con la ejecución de la propuesta y un acercamiento a la justificación del tipo de aporte.

Área problemática

Para abordar teórica y conceptualmente el tópico de transformación de lenguaje natural a lenguaje controlado en la educción de requisitos, es necesario partir de la definición de un objeto real de investigación. El objeto real se representa con la descripción de una necesidad de un interesado, expresada en *lenguaje natural* dentro de un documento técnico, y su traducción a un *lenguaje controlado* que especifica los requisitos (puede ser el lenguaje UN-Lencep; Zapata, 2007).

A partir del objeto real, se puede delimitar el objeto de estudio considerando dos conceptos básicos: lenguaje natural y lenguaje controlado. Estos conceptos se aplican el marco de la educción de requisitos utilizando técnicas de procesamiento del lenguaje natural y de la lingüística computacional.

Educción de requisitos

Es una de las primeras fases de la *ingeniería de requisitos* para el desarrollo de software, cuyo objetivo principal es descubrir todos los requisitos que el futuro software necesita satisfacer para que se considere de calidad. Para lograr este objetivo, se deben llevar a cabo, de manera iterativa e incremental, dos actividades primordiales: *educción de requisitos* y *análisis de requisitos*, involucrando un lenguaje natural y un lenguaje de modelado, respectivamente (Li *et al.*, 2003).

La fase de educción de requisitos tiene diferentes actividades, que incluyen: entendimiento del dominio de aplicación, captura y clasificación de requisitos, establecimiento de prioridades, resolución de conflictos y negociación de los requisitos del sistema (Robertson y Robertson, 2006). La educción de requisitos se relaciona, principalmente, con la acción de comunicación analista-interesado, la cual busca recuperar la información esencial y relevante acerca del dominio, obtener la base de los requisitos y extraerla de los interesados (cliente, usuario-final, experto del dominio, etc.).

Lenguaje natural y lenguaje controlado

Serrano (2005), citando a Ferdinand de Saussure, expresa que el lenguaje se concibe como el complemento de dos entidades: *lengua* y *habla*. El lenguaje es propiedad social, no individual, como la totalidad de los sistemas lingüísticos que emplean los miembros de una comunidad, es decir, es un sistema de signos. Así, de acuerdo con Vernengo (1996), el lenguaje se puede entender como un conjunto de oraciones gramaticalmente bien formadas conforme a reglas fonéticas, léxicas, sintácticas y semánticas correspondientes a un lenguaje natural cualquiera. En su estado normal, el lenguaje natural utiliza elementos gramaticales, como: sustantivo, verbo, adjetivo, pronombre, conjunción, preposición, adverbio y artículo.

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

Por lenguaje natural se entiende la lengua utilizada normalmente en una comunidad de individuos para la comunicación de estos entre sí (Tendales, 2004). El lenguaje natural se caracteriza por su enorme capacidad y su riqueza comunicativa, su flexibilidad y la posibilidad de jugar con las palabras y con las expresiones, produciendo metáforas y ambigüedades. De lo anterior se deduce que, si bien el lenguaje natural es un instrumento idóneo para ciertos propósitos, no lo es igualmente para áreas científicas o ingenieriles donde se requiere un máximo de exactitud y precisión (Li *et al.*, 2003).

Según Berry (2003), la gran mayoría de requisitos se escribe en lenguaje natural. Para el analista es muy importante identificar los conceptos y relaciones entre conceptos que emplea el interesado, que constituyen la base del lenguaje común que deben entender el analista y el experto. En general, según Li *et al.*

(2003), el lenguaje natural es altamente informal por naturaleza, lo que implica mayor intervención y transformación, para las tareas posteriores de análisis y diseño de un producto software.

Un lenguaje controlado (LC), según Wojcik y Hoard (1995), es un subconjunto del lenguaje natural con sintaxis, semántica o terminología restringidas. 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.

Procesamiento de lenguaje natural

La transformación, cuando se habla del procesamiento del lenguaje natural, se refiere a la traducción de la versión de un texto desde una lengua natural a otra (Moreiro, 1992). Para procesar el lenguaje natural se requiere transformar el texto en una representación semántica apta para razonar, tomar decisiones y ejecutar ciertas tareas (Lourdes, 2006). Esta representación se consigue por medio del proceso de *parsing* o construcción de un árbol de análisis a partir de una gramática (Gavaldá, 2011). Si la gramática es sintáctica, por medio de un árbol de análisis se genera información sobre las categorías gramaticales de las palabras y la función sintáctica asociada (por ejemplo, la identificación del sujeto, el verbo, el predicado, los complementos, etc.). Mientras tanto, si la gramática es semántica, el árbol de análisis ya es bastante próximo a la representación lógica que permite el razonamiento y la ejecución.

Lingüística computacional

Diferentes disciplinas dentro del campo de la *lingüística* estudian el *lenguaje*. A su vez, este campo se ocupa de todos los hechos y fenómenos relacionados con el lenguaje natural. El objetivo de la *lingüística* es producir modelos que se aproximen al comportamiento humano en sus tareas básicas: leer, escribir, escuchar y hablar. Este campo, según Castro *et al.* (2009), se enfoca en el estudio de los signos lingüísticos e incluye la semántica, la sintaxis y la pragmática. 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).

El término *lingüística computacional* (en inglés *computational linguistics*) se refiere al campo interdisciplinario entre 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 los 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.

Problema de investigación

En el marco de la ingeniería de requisitos, para iniciar el proceso de educción, se requiere descubrir y obtener el máximo de información para el conocimiento de un contexto en cuestión. El discurso contiene esta información. Una vez se consolida un discurso que describe el dominio del problema, el analista representa, mediante un modelo, el ámbito del dominio y su solución; normalmente, se utiliza un modelo conceptual. Para desarrollar un modelo conceptual, el analista o diseñador debe identificar ciertos elementos conceptuales, identificar las relaciones entre ellos y entender esta relación, para luego representar esos elementos en un lenguaje de modelado (Gangopadhyay, 2001). Este proceso se puede visualizar como una 'traducción' de un lenguaje base a otro diferente, de manera tal que el traductor reconozca y entienda símbolos expresados en un lenguaje natural de un universo de discurso, en un conjunto de símbolos definidos en un lexicón de un lenguaje de modelado (Castro *et al.*, 2009).

Es en este proceso de traducción donde el analista, luego de capturar las necesidades y expectativas del interesado, las representa en modelos técnicos. Los interesados, por su parte, validan los requisitos capturados y plasmados en dichos modelos (que suelen ser en su mayoría gráficos), aunque no los comprenden suficientemente, porque se describen en un lenguaje técnico que se aleja del natural. Esta comunicación entre los participantes (interesado-analista) se torna más débil aún por la diferencia de formación y experiencia entre ellos (Zapata & Villa, 2008), lo que genera problemas en cuanto a la correcta validación de los requisitos, el alcance de los modelos conceptuales generados y, finalmente, los altos costos de ejecución del proceso de educción.

Por otro lado, tradicionalmente, la obtención de requisitos parte de la aplicación de técnicas de captura,

como entrevistas y diseño de aplicaciones conjuntas (Christel & Kang, 1992), u otras técnicas enfocadas hacia el análisis de escenarios, como las que describen Zapata *et al.* (2007). No es muy común educir requisitos a partir de otro tipo de fuentes, como la documentación técnica, la cual incluye información en forma de manuales de procedimientos, reglamentos y estatutos de organización, etc. Esta educción permitiría principalmente: una comprensión y descripción detallada de la propia organización y del papel que representa el sistema en este contexto (Leite, 1987), la comprensión del dominio del interesado, el diseño posterior de entrevistas, la aplicación de técnicas de análisis de requisitos y la generación de modelos iniciales del dominio del problema.

A partir de las técnicas de captura de requisitos utilizadas de forma tradicional, la intervención del interesado se suele estimar incorrectamente, pues, la mayoría de las veces, se realiza en forma de diálogos y entrevistas (Leite, 1987). En este proceso se pierde tiempo, secuencia, coherencia y concisión, dado que los interesados tienden a dilatar sus intervenciones y la entrega de información, lo que, como ya se indicó, acarrea mayores tiempos en la educción y mayores costos. El compendio de información obtenida y las descripciones del dominio de aplicación, que son resultado del trabajo con el interesado, tienen los problemas propios del lenguaje natural: mucha información, uso indiscriminado de sinónimos y ambigüedades, etc.

A pesar de los acercamientos propuestos en la literatura para reducir la brecha entre los universos de discurso del interesado y del analista, todavía se exige máxima intervención del analista en el proceso de educción, en la conversión entre la descripción de requisitos y los modelos de diseño y en el método que guíe este proceso. Es necesario lograr acercamientos entre los lenguajes controlados que existen para especificar los requisitos y el lenguaje natural del interesado, el cual se puede traducir directamente a partir de documentación técnica y así conducir a las etapas posteriores del proceso de desarrollo del producto software, que se realiza actualmente de forma automática, como propone Zapata (2007).

A partir de la descripción problemática anterior, se plantea la siguiente pregunta de investigación: **¿***Cómo especificar un proceso de transformación automático de lenguaje natural a lenguaje controlado, a partir de documentación técnica, para la educción de requisitos en el diseño de un producto de software*?

Conclusiones y resultados esperados

En la literatura no se cuenta con un modelo descriptivo que represente el proceso de transformación de las necesidades y expectativas del interesado, expresadas en lenguaje natural, en requisitos expresados en un lenguaje controlado, en la fase de educción de requisitos a partir de documentación técnica. Es necesaria una formalización de dicho proceso, a partir de las teorías que ofrecen la lingüística computacional y el procesamiento de lenguaje natural, lo que podría derivar en la propuesta de nuevos conceptos o la inclusión de conceptos de otras disciplinas del procesamiento de lenguaje, en la ingeniería de requisitos.

Es necesario realizar aportes que permitan, entre otros, generar los siguientes resultados esperados:

- Facilitar la tarea de modelado del analista, a partir de información capturada de documentación técnica que se pueda representar en un lenguaje técnico o modelo técnico, aplicando un método establecido.
- Mejorar la comprensión de los interesados sobre los modelos y esquemas conceptuales que diseñan los analistas, por medio de un lenguaje cercano al natural.
- Proveer técnicas y formalismos que permitan trasladar descripciones y conocimiento de la organización, hacia los modelos cercanos al proceso de análisis de requisitos.
- Definir un marco conceptual respecto de las variables que intervienen en el proceso de transformación de lenguaje natural a un lenguaje controlado, a partir de documentación técnica, en la educción de requisitos.
- Extender, por medio de un formalismo o procedimiento, el proceso de transformación de lenguaje natural a lenguaje controlado.
- Mostrar cómo ciertas propiedades, teorías o herramientas utilizadas en la lingüística, se pueden utilizar en el marco del procesamiento del lenguaje natural, para mejorar el proceso de transformación de lenguaje a partir de documentación técnica, en la educción de requisitos.

Agradecimientos

Este trabajo se enmarca dentro de los resultados obtenidos en el proyecto de investigación 'Revisión de Literatura en Transformación de Lenguaje Natural a Lenguaje Controlado en la Educción de Requisitos', cofinanciado entre la Universidad Nacional de Colombia, Sede Medellín, y la Universidad de Medellín, Colombia.

Referencias

Bolton, D., Jones, S., Till, D., Furber, D. & S. Green (1994). Using domain knowledge in requirements capture and formal specification construction. *Requirements Engineering: Social and Technical Issues*, Academic Press, 2ª. ed., pp. 141-162.

Castro, L., Baiao, F. & Guizzardi, G. (2009). A survey on conceptual modeling from a linguistic point of view. *Relatórios t*écnicos do departamento de informática aplicada *da Unirio,* N° 0019/2009, pp. 3-12.

Clegg, A. (2008). *Computational-linguistic approaches to biological text mining*. Tesis de PhD. Londres: Escuela de Cristalografía, University of London.

Cunningham, H. (2000). *Software architecture for language engineering*. Tesis de PhD. Reino Unido: Departamento de ciencias de la computación, University of Sheffield.

Cheng, B. & Atlee, J. (2007). *Research directions in requirements engineering*. Proceedings of future of software engineering (FOSE'07), IEEE Computer Society, USA.

Christel, M. & Kang, K. (1992). *Issues in requirements elicitation*. Technical report CMU/SEI-92-TR-012 ESC-TR-92-012. USA: Software Engineering Institute.

Gangopadhyay, A. (2001). Conceptual modeling from natural language functional specifications. *Artificial Intelligence in Engineering*, Vol. 15, No. 2, pp. 207-218.

Gavaldá, M. (2011). La investigación en tecnologías de la lengua. Research in language technology. http://quark.prbb.org/19/019021.htm [Consultado el 15 de mayo de 2011].

Haller, J. & Schütz, J. (2001). *CLAT: Controlled language authoring technology.* Proceedings of the 19th annual international conference on computer documentation, Santa Fe NM.

Kof, L. (2004). *Natural language processing for requirements engineering: applicability to large requirements documents*. Alemania: Fakultät für Informatik, Technische Universität München.

Leite, J. (1987). *A survey on requirements analysis*. Advanced software engineering project technical report RTP071. EE.UU.: Department of Information and Computer Science, University of California.

Lourdes, A. (2006). Procesamiento de lenguaje natural. Disponible http://tabasco.torreingenieria.unam.mx/gch/PLN/cap1.pdf [Consultado el 10 de mayo de 2011].

Moreiro, J. (1992). *Perspectiva documental del procesamiento de lenguaje natural*. Memorias Congreso SEPLN VIII, Universidad Carlos III, Madrid.
Serrano, W. (2005). ¿Qué constituye a los lenguajes natural y matemático? *Sapiens: Revista Universitaria de Investigación*, Vol. 6 No. 001, pp. 47-59.
Tendales (s.f.). Lógica simbólica. Lógica proposicional. http://blog.educastur.es/tendales/files/2009/12/logica-teoria2.pdf
[Consultado el 25 de mayo de 2011].

Vernengo, R. (1996). El discurso del derecho y el lenguaje normativo. *Isonomía*, No. 4, pp. 87-95.

Wojcik, R. & Hoard, J. Controlled languages in industry. http://www.cslu.ogi.edu/HLTsurvey/ch7node8.html [Consultado el 22 de mayo de 2011].

Zapata, C.M. (2007). *Definición de un esquema preconceptual para la obtención automática de esquemas conceptuales de UML*. Tesis doctoral doctorado en ingeniería. Colombia: Universidad Nacional de Colombia Sede Medellín.

Zapata, C.M., Palacio, C. & Olaya, N. (2007). UNC-ANALISTA: Hacia la captura de un corpus de requisitos a partir de la aplicación del experimento Mago de Oz. *Revista EIA*, N. 7, pp. 25-40.

Zapata, C.M. & Villa, F. A. (2008). La gramática básica de UN-Lencep expresada en HPSG. *Avances en Sistemas e Informática*, Vol.5 No.1, edición especial, pp. 81-92.