Adecuación de la forma de la construcción de glosarios al estilo del discurso

Renata S. Guatelli[1], Gladys N. Kaplan[1], Jorge H. Doorn[2,3]

[1]Departamento de Ingeniería e Investigaciones Tecnológicas – Universidad Nacional de La Matanza
[2]Escuela de Informática, Universidad Nacional del Oeste
[3]Departamento de Ingeniería, Universidad Nacional de Tres de Febrero
{rguatelli; gkaplan}@unlam.edu.ar; jdoorn@uno.edu.ar;

RESUMEN

Es conocida la importancia que tiene la construcción de un glosario en los procesos de requisitos. La experiencia recogida a través de numerosos casos reales ha mostrado que la construcción del mismo es simultáneamente laboriosa y que sus resultados son pocos confiables. Es razonable suponer que el uso de estrategias de procesamiento de lenguaje natural puede contribuir a atemperar ambas dificultades. La construcción de glosarios difiere de la minería de textos clásica en el sentido que se tiene un cierto conocimiento previo de aquello que se busca. Sin embargo, tanto debido a las peculiaridades de la fuente de información como a las características del Universo de Discurso, los estilos de las narrativas ofrecen una sensible dispersión. En el presente proyecto se planifica utilizar un enfoque de ingeniería inversa, utilizando fuentes de información confiables y glosarios ya construidos por seres humanos para estudiar el contexto concreto de uso de los símbolos incluidos en el glosario en la fuente de información. De este estudio se espera deducir reglas que permitan detectar símbolos no descubiertos a partir de la combinación del contexto de uso y del estilo del discurso.

Palabras clave: *Lenguaje natural, entrevista, transcripción, LEL, dispersión de información.*

CONTEXTO

La propuesta que se presenta es parte del proyecto de investigación “Aspectos No Funcionales de los Procesos de Requisitos” del Departamento de Ingeniería de Investigaciones Tecnológicas (DIIT) de la Universidad Nacional de La Matanza (UNLaM).

1.0 INTRODUCCIÓN

No todos los proyectos de software concluyen con éxito [1] [2] [3] [4]. Diversos estudios llevados a cabo para analizar este tema argumentan que el problema principal encontrado son requisitos inadecuados, mal comprendidos, incompletos y volátiles. Para poder definir los requisitos primero se debe adquirir conocimiento sobre el Universo de Discurso (UdeD), capturar las demandas, necesidades y problemas presentes en él. Esta información es analizada por la ingeniería de requisitos (IR), con el fin de generar la Especificación de Requisitos de Software (ERS) que contiene los servicios que el sistema de software debe satisfacer para mejorar los procesos del negocio. Para Loucopoulos [5] la IR se compone de actividades que permiten comprender las necesidades de los clientes-usuarios y traducir dichas necesidades en declaraciones precisas y sin ambigüedades que posteriormente se pueden utilizar en el proceso de desarrollo del nuevo sistema de software. Una ERS adecuada, es la base de las actividades de Gestión de Proyecto relacionadas con el presupuesto y el cronograma, influyendo sobre la calidad del mismo.

La baja calidad de los productos de software desarrollados, ERS deficientes, elevados costos de corrección y mantenimiento, o el completo fracaso del proyecto, podrían evitarse o al menos mitigarse si hubiera una mayor preocupación por la rigurosidad durante las actividades relacionadas con la IR [6].

Los problemas mencionados, en su mayoría, tienen su origen en la utilización de modelos que los clientes-usuarios no pueden comprender. Mejorar las estrategias de comunicación entre los involucrados es de gran importancia para identificar, validar y verificar de manera correcta y oportuna dichos modelos [7].

La etapa inicial de la IR, implica la obtención o transferencia de la mayor parte del conocimiento desde el UdeD. Plasmar ese conocimiento en modelos basados en lenguaje natural (LN), además de mejorar la comunicación entre los participantes, asegura que su validación sea más segura y acertada. Los clientes-usuarios comprenden mucho mejor las descripciones basadas en LN en lugar de aquellas basadas en esquemas técnicos. En especial si este lenguaje es lo más cercano posible a la jerga utilizada por ellos.

El LN tiene un gran poder expresivo, pero es polisémico ya que contiene ambigüedad, ironías, expresiones típicas y muchas otras características, que generan problemas de interpretación, dado que su significado depende de los puntos de vista y del modelo mental de los interlocutores. A pesar de sus inconvenientes, el uso del LN mejora significativamente la transferencia de conocimiento, asegurando una mejor comprensión de los modelos [8].

En el campo la IR existen diferentes modelos basadas en LN, algunos son:

- Casos de Uso [9], son descripciones narrativas de la interacción entre un actor y el sistema.
- Glosarios [10] [11] contienen palabras y/o expresiones, comentadas o explicadas. Se los crea con diferentes fines, por ej. aclarar el significado de conceptos del dominio de la aplicación, unificar la terminología empleada en los diferentes modelos, mejorar la comunicación entre los involucrados; se pueden centrar en la terminología de los clientes - usuarios, o en la de los documentos.
- El LEL [12], es un glosario que describe el vocabulario de la aplicación, sin necesidad de comprender la funcionalidad del proceso del negocio. Las palabras o frases que contiene son llamados símbolos. Cada símbolo se identifica con un nombre (o más de uno en caso de sinónimos). Se detalla indicando la noción (denotación, define su significado) e impacto (connotación, identifica la relación del símbolo que se está describiendo con los demás símbolos del léxico). Noción e impacto se deben describir teniendo en cuenta el "principio de circularidad" (maximizar el uso de símbolos pertenecientes al LEL) y el uso de "vocabulario mínimo" (acotar el uso de lenguaje externo al dominio de la aplicación). Generalmente los símbolos se clasifican en sujeto, objeto, verbo y estado, de acuerdo a su uso en el dominio. Pueden crearse clasificaciones especiales.
- Escenarios [13], son descripciones de las situaciones que ocurren en el contexto. Pueden representar situaciones actuales o la planificación de situaciones futuras.
- Historias de Usuarios [14], consisten en funcionalidades descriptas por el

propio usuario. Usualmente responden al siguiente formato: i) Quién requiere la funcionalidad, ii) Cuál es la funcionalidad y iii) Por qué esa funcionalidad es necesaria (opcional).

Sin importar el modelo que se utilice, todos están afectados por los beneficios e inconvenientes del LN. Algunos utilizan el LN sin restricciones, otros intentan atemperar los posibles inconvenientes.

Dentro de las distintas fuentes de información, las personas son las más apropiadas, pero también son las que requieren un tratamiento más elaborado. Por lo tanto, la entrevista es la técnica de elicitación más utilizada [15] [16]. Estos datos han sido avalados por varios estudios [17] [18] los cuales confirman que se suelen privilegiar la entrevista por sobre otras técnicas de elicitación. A partir de las entrevistas se pueden generar una serie de productos intermedios, tales como minutas, anotaciones y transcripciones. Estos productos intermedios son algunas de las fuentes de información que se utilizan en los procesos de la IR.

Se debe prestar una debida atención a que cada persona consultada tiene su propio estilo de discurso y este puede ser narrativo, expositivo o argumentativo. Además, este discurso puede estar orientado a describir el proceso que se está analizando, el producto final o hacer hincapié en los componentes necesarios y sus interrelaciones.

Las narrativas de los entrevistados, especialmente en las entrevistas no estructuradas, suelen diferir mucho unas de otras, tanto por las propiedades del UdeD como por el punto de vista del entrevistado. Por ejemplo, en un sistema de naturaleza hospitalaria las narrativas estarán fuertemente influenciadas por los protocolos médicos por lo que serán orientados a los procesos. En cambio, es frecuente encontrar narrativas que describen entidades y relaciones entre las mismas, especialmente en organizaciones administrativas.

2.0 LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Este proyecto sigue la línea de investigación correspondiente al proceso de requisitos [20] basado en modelos en LN, orientado a atender las necesidades del cliente-usuario.

Se utilizarán técnicas de procesamiento de lenguaje natural, aplicadas sobre el modelo LEL, para mejorar la calidad del mismo, procurando reducir la notoria incompletitud observada en estudios previos [19] [20] [21] [22]. Usar estrategias de procesamiento de LN implica el uso de textos, por lo que se utilizará en todos los casos las transcripciones de las entrevistas en total concordancia con la enorme experiencia de las ciencias sociales que ampliamente las promueven [23] [24] [25]. Fundamentalmente en las transcripciones de las entrevistas es donde mejor se puede analizar el estilo discursivo para deducir nuevas reglas.

La técnica planificada consiste en una suerte de ingeniería inversa ya que se tomará un caso ya estudiado, y se analizarán todas las ocurrencias de los símbolos del LEL detectados por otros autores, y se estudiarán los conectores o frases que ligan los pares de símbolos en el texto.

En este sentido, se supone que los conectores lingüísticos o palabras auxiliares que ligan los símbolos del LEL en el discurso son fuertemente dependientes del estilo de la narrativa. Por ejemplo, es esperable que la frase “es un”, “es el”, “es parte de”, “está compuesto por” o similares estén presentes en un discurso centrado en entidades y ausentes en un discurso centrado en procesos.

Este estudio se planifica en el contexto de una calificación previa del discurso e intenta detectar las frases que vinculan los símbolos del

LEL. Estas frases serán luego útiles para mejorar la detección de símbolos en futuros proyectos. Obviamente, se panifica replicar el mismo estudio en varios casos para intentar determinar el grado de dependencia de estos conectores con el estilo del discurso utilizado y la variabilidad dentro de un mismo estilo.

Esta estrategia se basa en la hipótesis que en todo proyecto siempre existen unos pocos términos que son muy evidentes para el ingeniero de requisitos, los que serán incluidos en el LEL sin lugar a dudas. Es así que se planifica que la estrategia final resultante, utilice estos pocos símbolos iniciales como núcleo en la búsqueda semiautomática de nuevos símbolos. Es justamente en esa búsqueda de nuevos símbolos en la que el conocimiento de los conectores más usuales, posiblemente enriquezca y facilite la misma. Se espera que esto ayude a disminuir la incompletitud del LEL, permitiendo mejorar las heurísticas para colaborar en la detección de símbolos no triviales.

3.0 RESULTADOS ESPERADOS

En experiencias preliminares se han estudiados casos reales utilizando el mismo patrón de trabajo que se planifica utilizar sistemáticamente en el presente proyecto. Los resultados obtenidos son promisorios en el sentido que se han descubierto regularidades que de ser confirmadas permitirían guiar el procesamiento del LN con reglas basadas en el estilo del discurso.

En estos trabajos preliminares se transcribieron audios de entrevistas. En estos documentos se marcaron los símbolos que figuraban en el LEL, estudiando las palabras o frases que actuaban como conectores entre pares de símbolos, encontrándose que el número de conectores utilizados en la narrativa es relativamente reducido.

Como resultado principal del presente proyecto se espera detectar en los diferentes estilos discursivos cuáles son los conectores más frecuentes entre pares de símbolos del LEL. Estos se utilizarán luego para construir una herramienta que utilice las ocurrencias de los mismos para sugerir posibles nuevos términos a ser incluidos en el LEL.

4.0 FORMACIÓN DE RECURSOS HUMANOS

La línea de investigación presentada colabora en la tesis doctoral de la Mg. Gladys Kaplan y es parte directa de la tesis de maestría de la Lic. Renata Guatelli.

5.0 BIBLIOGRAFÍA

[1] Gibbs, W. W. (1994). Software's chronic crisis. Scientific American, 271(3), 86-95.

[2] Finkelstein, A., & Dowell, J. (1996). A comedy of errors: the London Ambulance Service case study. In Proceedings of the 8th International Workshop on Software Specification and Design, pp. 2-4.

[3] Lindstrom, D. R. (1993). Five ways to destroy a development project (software development). IEEE Software, 10(5), pp.55-58.

[4] El Emam, K., & Koru, A. G. (2008). A replicated survey of IT software project failures. IEEE software, 25(5), pp.84-90.

[5] Loucopoulos, P., & Karakostas, V. (1995). System requirements engineering. McGraw-Hill, Inc.

[6] de Almeida Ferreira, D., & da Silva, A. R. (2009). A controlled natural language approach for integrating requirements and model-driven engineering. In Fourth International Conference on Software Engineering Advances, pp. 518-523.

[7] Boehm, B. W. (1984). Verifying and validating software requirements and design specifications. IEEE software, (1), pp. 75-88.

[8] Jackson, M. (1995). Software requirements & specifications: a lexicon of practice, principles and prejudices. ACM Press/Addison-Wesley Publishing Co.

[9] Jacobson, I. (1993). Object-oriented software engineering: a use case driven approach. Pearson Education.

[10] Weidenhaupt K., Pohl K., Jarke M., Haumer, P (1998) Scenarios in System Development: Current Practice, IEEE Software, pp 34-45.

[11] Robertson S. and Robertson J. (2006) Mastering the Requirements Process, 2nd Ed, AddisonWesley.

[12] Leite, J. D. P., & Franco, A. P. M. (1993, January). A strategy for conceptual model acquisition. In Proceedings of the IEEE International Symposium on Requirements Engineering, pp. 243-246.

[13] Carroll, J. M. (Ed.). (1995). Scenario-based design: envisioning work and technology in system development. John Wiley & Sons.

[14] Beck, K. (2000). Extreme programming explained: embrace change. Addison-Wesley.

[15] Pan, D., Zhu, D., & Johnson, K. (2001). Requirements Engineering Techniques. Internal Report. Department of Computer Science. University of Calgary. Canada.

[16] Bourque, P., & Fairley, R. E. (2014). Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0. IEEE Computer Society Press.

[17] Antonelli, L., & Oliveros, A. (2002). Fuentes Utilizadas por desarrolladores de Software en Argentina para Elicitar Requerimientos. In fifth Workshop on Requirements Engineering, pp. 106-116.

[18] Oliveros, A., & Antonelli, R. L. (2015). Técnicas de elicitación de requerimientos. In XXI Congreso Argentino de Ciencias de la Computación, pp.546-555

[19] Ridao, M., & Doorn, J. H. (2006). Estimación de Completitud en Modelos de Requisitos Basados en LN. In 9th Workshop on Requirements Engineering, pp.146-152.

[20] Litvak, C. S., Hadad, G. D., & Doorn, J. H. (2012). Un abordaje al problema de completitud en requisitos de software. In XVIII Congreso Argentino de Ciencias de la Computación, pp. 827-836.

[21] Litvak, C. S., Hadad, G. D., & Doorn, J. H. (2013). Mejoras semánticas para estimar la Completitud de Modelos en Lenguaje Natural. In 1er Congreso Nacional de Ingeniería Informática / Sistemas de Información, (p.9) http://conaiisi.frc.utn.edu.ar/PDFsParaPublicar/1/schedConfs/7/7-477-1-DR.pdf (consultado el 26/03/2020)

[22] Litvak, C. S., Hadad, G. D. S., & Doorn, J. H. (2016). Procesamiento de lenguaje natural para estudiar completitud de requisitos. In XVIII Workshop de Investigadores en Ciencias de la Computación, pp. 498-502

[23] Valles, M. S. (2007). Entrevistas cualitativas (Vol. 32). CIS.

[24] Kvale, S. (2011). Las entrevistas en investigación cualitativa. Ediciones Morata.

[25] Díaz-Bravo, L., Torruco-García, U., Martínez-Hernández, M., & Varela-Ruiz, M. (2013). La entrevista, recurso flexible y dinámico. Investigación en educación médica, 2(7), pp. 162-167.