Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 219-222 recibido 29-04-2013 revisado 16-06-2013 aceptado 21-06-2013

LEGOLANG: Técnicas de deconstrucción aplicadas a las Tecnologías del Lenguaje Humano

LEGOLANG: Deconstruction Techniques applied to Human Language Technologies

P. Martínez-Barco, A. Ferrández, D. Tomás, E. Lloret, E. Saquete, F. Llopis, J. Peral, M. Palomar, J.M. Gómez
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{patricio,antonio,dtomas,elloret,stela,llopis,jperal,mpalomar,jmgomez}@dlsi.ua.es
M.T. Romá-Ferri
Departamento de Enfermería
Universidad de Alicante
mtr.ferri@ua.es

Resumen: El objetivo de este proyecto se basa en la necesidad de replantearse la filosofía clásica del TLH para adecuarse tanto a las fuentes disponibles actualmente (datos no estructurados con multi-modalidad, multi-lingualidad y diferentes grados de formalidad) como a las necesidades reales de los usuarios finales. Para conseguir este objetivo es necesario integrar tanto la comprensión como la generación del lenguaje humano en un modelo único (modelo LEGOLANG) basado en técnicas de deconstrucción de la lengua, independiente de su aplicación final y de la variante de lenguaje humano elegida para expresar el conocimiento.
Palabras clave: tecnologías del lenguaje humano (TLH), comprensión del lenguaje, generación del lenguaje, desconstrucción del lenguaje

Abstract: The main objective of this project is based on the need to reconsider the clasical HLT philosophy to adapt it, not only to the currently available resources (unstructured data with multimodality, multilinguality and different levels of formality) but also to the real needs of the final users. In order to reach this objective it is necessary to include the understanding as well as the generation of human language in a unique model (LEGOLANG model) based on language deconstruction techniques, independently of the final application and the human language variant chosen to express the knowledge.
Keywords: human language technologies (HTL), human language understanding, human language generation, language deconstruction

1 Introducción

Se conoce como generación del lenguaje natural (GLN) al proceso de construcción deliberada de texto en lenguaje natural con el fin de alcanzar capacidades comunicativas previamente especificadas (McDonald, 1987). Con este objetivo, la GLN se convierte en elemento indispensable para múltiples aplicaciones que derivan en fines más concretos como la construcción automática de informes estandarizados, la producción automática de resúmenes, la traducción automática, etc. De esta manera, y tomando como base la definición de Reiter & Dale (1997), podemos hablar de GLN como una línea de investigación en el ámbito de las Tecnologías del Lenguaje Humano (TLH), cuyo fin último es el de proporcionar un conjunto de herramientas y técnicas capaces de producir texto comprensible en lenguaje natural a partir de una representación no lingüística de la información, generalmente, desde bases de datos o fuentes de conocimiento.

Otra de las grandes líneas que emanan de las TLH es la compresión del lenguaje natural (CLN) que trata de extraer, de manera automática, el significado de un texto dado y obtener una representación estructurada del mismo para su uso posterior. Así, GLN y CLN podrían llegar a entenderse como grandes

ISSN 1135-5948

procesos de análisis simétricos. Sin embargo, la investigación tradicional de TLH ha disociado sus líneas en las dos grandes ramas citadas, GLN y CLN, dando lugar a un conjunto de aproximaciones diferentes, que si bien parten de hipótesis teóricas comunes, sus realizaciones finales distan en muchos casos de ser compatibles (Reiter & Dale, 2000).

Además, la nueva situación implica que los sistemas de GLN deben acometer la captura de información desde colecciones documentales no estructuradas, multilingües y multimodales, con escasas garantías de fiabilidad y diversos grados de formalidad, provenientes de fuentes tan dispersas y diversas como artículos periodísticos, informes técnicos, blogs, microblogs, wikis o redes sociales. Esta situación deriva en un problema aún sin resolver, y es que no existe un modelo único de comprensión y generación del lenguaje independiente de la aplicación.

2 *Propuesta*

Nuestra propuesta consiste en plantear una concepción del lenguaje humano totalmente novedosa en la que se descontextualizará el concepto de deconstrucción para redefinirlo en el marco de las Tecnologías del Lenguaje Humano, como un modelo que permitirá descomponer textos conocidos en un caos de unidades básicas de conocimiento (fase de comprensión del lenguaje) que, mediante la apropiada definición de nuevas estructuras, volverá a combinarse para proporcionar nuevos conocimientos (fase de generación del lenguaje). La deconstrucción, así entendida, nos permitirá modelar una nueva metodología para la generación de un lenguaje humano no tan centrado en la definición de estructuras gramaticales correctas sino de estructuras prácticas que muestren al receptor nuevos conocimientos ocultos en los documentos originales.

En consecuencia, este proyecto persigue tres metas fundamentales:

1. La definición de una unidad básica de conocimiento orientada al GLH a la que denominaremos L-Brick (Language Brick, o Ladrillo de Lenguaje).
2. El modelado del proceso de deconstrucción que, a partir de una colección documental, debe ser capaz de generar la representación del mismo en un sistema caótico de L-Bricks, definiendo el conjunto de recursos y técnicas útiles para dotar de contenido necesario a esas estructuras.
3. El rediseño de las tareas de los sistemas clásicos de GLN en función de los L-Bricks y de sus reglas de composición, de tal manera que permitan definir nuevas formas de comunicación del conocimiento, tomando como única base la información contenida en ellos.

En la Figura 1 se puede ver gráficamente la propuesta de este proyecto.

3 *Antecedentes*

Como se ha comentado previamente, la investigación clásica en Tecnologías del Lenguaje Humano ha tratado a las técnicas de generación de lenguaje natural (GLN) de manera aislada respecto a las de comprensión (CLN). En concreto, el estudio de la gran mayoría de trabajos iniciales de GLN (Bernardos, 2007) coincide en considerar una arquitectura común cuyo origen es siempre una representación computacional de la información, y un posterior procesamiento de la misma basado en dos grandes niveles (Hovy, 2000): a) determinación del qué decir y b) determinación del cómo decirlo. Para abordar estos dos niveles, Reiter E. (1994) propone tres fases diferentes: Macroplanning (para el qué decir), Microplanning (para el cómo decirlo), y Realización (para decirlo).

Además, en los últimos años ha surgido un

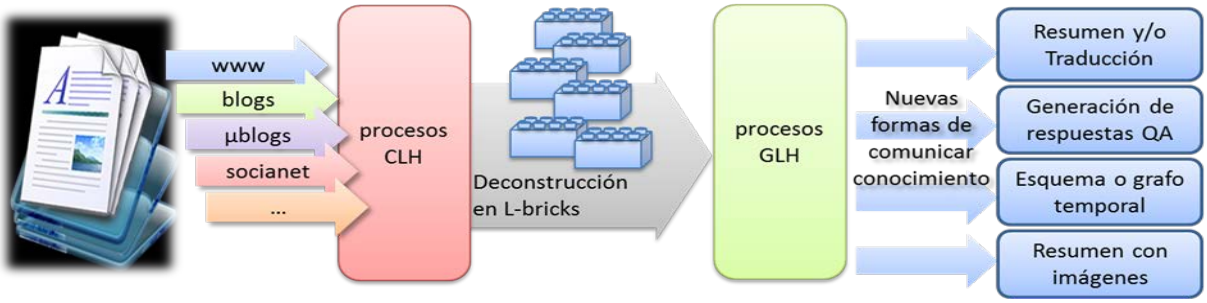

Figura 1: Propuesta del proyecto

creciente interés por abordar el problema de la generación del lenguaje no únicamente desde la componente "natural" (GLN, generación de lenguaje textual, formal y sintácticamente correcto), sino en general, desde la vertiente "humana" (GLH, generación de cualquier tipo de lenguaje para comunicación entre humanos). Cabe destacar, por poner algunos ejemplos, conferencias específicas como Generation Instructions in Virtual Environments (GIVE, 2011), Workshop on Multimodal Output Generation (MOG, 2011) y Generation Challenges (GC, 2011), donde se han presentado múltiples aproximaciones centradas en esta idea.

4 Metodología y plan de trabajo

La ejecución del proyecto se ha estructurado en las capas que se muestran en la Figura 2.

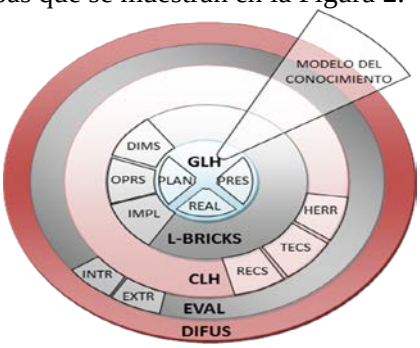

Figura 2. Estructuración modular del proyecto

La capa central, GLH (generación del lenguaje humano) es la motivación principal del proyecto, es el fin a alcanzar, conseguir generar lenguaje. Por encima de ella, la capa L-BRICKS (deconstrucción en unidades básicas de conocimiento) es una representación intermedia básica para poder construir la GLH. A su vez, la capa de CLH (comprensión del lenguaje humano) prestará sus servicios para dotar de contenido a la capa L-BRICKS. La capa EVAL (evaluación) recoge todas las actividades que nos permitirán conocer la validez de las propuestas de las capas internas. Finalmente, la capa más externa corresponde a las actividades de DIFUS (difusión) representando la piel del proyecto, es decir, lo que se va a dar a conocer de todas las actividades internas.

Por otra parte, el plano transversal (modelo del conocimiento) cruza las fronteras de todas las capas con el fin de identificar los hilos conductores comunes a todas aquellas tareas relacionadas con una misma motivación. Las actividades transversales son redundantes a las anteriores, pero sirven para dar coherencia al modelo desde una perspectiva diferente.

A continuación vamos a especificar un poco más en detalle las actividades y los hitos de cada una de estas capas.

Capa CLH: Comprensión del lenguaje humano

En esta capa se analizarán, recopilarán, adaptarán e integrarán todos los recursos, técnicas y herramientas necesarias para transformar la información obtenida desde diferentes fuentes en conocimiento útil que posteriormente se almacenará en las unidades básicas de conocimiento a través de tres actividades: a) *CLH.RECS: Recursos*, cuyo objetivo es la obtención y puesta a disposición del proyecto del conjunto de recursos necesarios para las tareas de comprensión del lenguaje, b) *CLH.TECS: Técnicas*, que se encargará de la recopilación e investigación del conjunto de técnicas necesarias para las tareas de comprensión del lenguaje, y c) *CLH.HERR: Herramientas*, cuyo hito es la obtención, implementación e integración del conjunto de herramientas necesarias para las tareas de comprensión del lenguaje.

Capa L-BRICKS: Deconstrucción del lenguaje en unidades básicas de conocimiento

En esta capa se tratarán las actividades relacionadas con la definición, estructuración e inserción de datos en estas unidades, denominadas L-Brick (language brick: ladrillo del lenguaje) por el paralelismo generado con las unidades de los juegos infantiles para construcción en bloques: a) *L-BRICKS.DIMS: Dimensiones*, en la que se definirá la estructura multidimensional del ladrillo, b) *L-BRICKS.OPRS: Operaciones*, cuyo objetivo es la planificación del conjunto de operaciones posibles en la unidad L-Brick, y c) *L-BRICKS.IMPL: Implementación*, cuya finalidad es la implementación computacional de la estructura, operaciones y almacenamiento del L-Brick.

Capa GLH: Generación del lenguaje humano

El objetivo de esta capa es la generación de lenguaje humano a partir de los L-Bricks, que son nuestras unidades básicas de conocimiento.

Por tanto, en esta capa, nos centraremos principalmente en el análisis de técnicas y herramientas para poder comunicar la información contenida en los L-Bricks, que se corresponde con la etapa de realización del proceso de GLH. Las actividades de esta capa son las siguientes: a) *GLH.PLAN: Planificación*, cuyo hito es la obtención de técnicas para planificar la presentación del conocimiento, b) *GLH.REAL: Realización*, cuya finalidad es la obtención de técnicas para la realización del conocimiento contenido en los L-Bricks, y c) *GLH.PRES: Presentación*, cuyo objetivo es la definición del modelo de presentación del conocimiento del L-Brick.

Capa EVAL: Evaluación

La capa EVAL contempla las actividades necesarias para la realización de la evaluación del proyecto en dos niveles: a*) EVAL.INTR: Evaluación intrínseca*, que pretende llevar a cabo el análisis y definición de una serie de métricas cualitativas y cuantitativas que permitan evaluar intrínsecamente el modelo de GLH definido, y b) *EVAL.EXTR: Evaluación extrínseca*, que plantea la definición de un escenario sobre el cual aplicar el modelo de GLH para realizar una evaluación extrínseca del mismo.

Capa DIFUS: Difusión de la investigación

El objetivo de esta actividad es la difusión de los resultados científicos y tecnológicos alcanzados durante el desarrollo del proyecto. Si bien la finalidad última es la definición del modelo basado en L-Brick y su explotación para la generación de lenguaje, existen numerosas tareas intermedias para llevar a cabo este objetivo que por sí mismas resultan de interés para la comunidad científica.

Plano transversal: Modelo de conocimiento

El plano transversal representa un conjunto de actividades relacionadas con la representación del modelo de conocimiento para cada uno de los niveles de análisis del lenguaje (léxico, sintáctico, semántico, pragmático), desde su comprensión en el origen hasta la generación en el destino, atravesando todas las capas de la arquitectura. Este plano actuará como mecanismo cruzado de cohesión entre las capas y de esta manera se garantiza que únicamente se realizarán esfuerzos en tareas de CLH cuando realmente tengan utilidad para tareas de GLH, y a su vez, todas las tareas de GLH obtengan el conocimiento necesario desde las tareas de CLH.

5 Agradecimientos

El proyecto LEGOLANG[i] está financiado por el Ministerio de Economía y Competitividad con número de referencia TIN2012-31224.

Bibliografía

Bernardos, S. 2007. ¿Qué es la generación de lenguaje natural? Una visión general sobre el proceso de generación. *Revista Iberoamericana de Inteligencia Artificial, 34*, 105-128.

GC. 2011. *Generation Challenges*. Obtenido de http://www.nltg.brighton.ac.uk/research/genchal10/

GIVE. 2011. *Generation Instructions in Virtual Environments*. Obtenido de http://www.give-challenge.org/research/

Hovy, E. 2000. Language Generation (article 86). En E. Reilly, A. Ralston, & D. Hemmendinger, *Encyclopedia of Computer Science*. London: McMillan.

McDonald, D. D. 1987. Natural Language Generation. En S. C. Shapiro, *Encyclopedia of Artificial Intelligence* (págs. 642-655). John Wiley and Sons.

MOG. 2011. *Workshop on Multimodal Output Generation* . Obtenido de http://www.mog-workshop.org/

Reiter, E. 1994. Has a Consensus NL Generation. *Proceedings of the 7th International Workshop on Natural Language Generation*, (págs. 163-170). Kennebunkport.

Reiter, E., & Dale, R. 1997. Building applied natural language generation systems. *Natural Language Engineering, 3*, 57-87.

Reiter, E., & Dale, R. 2000. *Building natural language generation systems*. Cambridge: Cambridge University Press.

[i] http://gplsi.dlsi.ua.es/proyectos/legolang/