Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 177-180 recibido 26-03-2015 revisado 25-4-2015 aceptado 12-05-2015

# Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario

## *Exploitation and Processing of Online Information for Annotating and Generating Texts Adapted to the User*

**Elena Lloret, Yoan Gutiérrez, Fernando S. Peregrino, José Manuel Gómez, Antonio Guillén, Fernando Llopis**
Universidad de Alicante
Carretera San Vicente del Raspeig s/n 03690, Alicante, España
{elloret,ygutierrez,fsperegrino,jmgomez,aguillen,llopis}@dlsi.ua.es

**Resumen:** La gran cantidad de información disponible en Internet está dificultando cada vez más que los usuarios puedan digerir toda esa información, siendo actualmente casi impensable sin la ayuda de herramientas basadas en las Tecnologías del Lenguaje Humano (TLH), como pueden ser los recuperadores de información o resumidores automáticos. El interés de este proyecto emergente (y por tanto, su objetivo principal) viene motivado precisamente por la necesidad de definir y crear un marco tecnológico basado en TLH, capaz de procesar y anotar semánticamente la información, así como permitir la generación de información de forma automática, flexibilizando el tipo de información a presentar y adaptándola a las necesidades de los usuarios. En este artículo se proporciona una visión general de este proyecto, centrándonos en la arquitectura propuesta y el estado actual del mismo.
**Palabras clave:** PLN, Ontología, Paquete semántico, Generación de textos

**Abstract:** The great amount of available online information is making increasingly more and more difficult that users can assimilate such as volume of information, being this almost inconceivable without using Human Language Technologies (HLT) tools, for instance, information retrieval systems or automatic summarisers. The interest of this emerging project (and therefore its main goal) is precisely motivated by the need to define and create a HLT-based technological framework, able to process and semantically annotate all this information, allowing also the automatic generation of information, and making the type of information to be presented more flexible by adapting it to the users' needs. This article provides an overview of this project, focusing on the proposed architecture and its current status.
**Keywords:** NLP, Ontology, Semantic package, Text Generation

## *1 Introducción y objetivo*

Actualmente, Internet cuenta con más de 2.400 millones de usuarios[1], dato que implica que más del 30 % de la población mundial está conectada. Además, desde la aparición de la Web 2.0 (o Web social), se han creado nuevos sitios Web dónde los usuarios juegan un papel más activo, a través de los que pueden participar, interactuar e intercambiar información con otros usuarios (por ejemplo, foros, blogs, redes sociales, microblogs, etc.).

Sin embargo, el principal inconveniente de toda esta gran cantidad de información disponible es la complejidad en lo que respecta a su procesamiento y tratamiento, sobre todo si el usuario desea obtener información con mayor o menor detalle acerca de un tema concreto. Dicha información se encuentra en distintas fuentes de información de distinta naturaleza y en distintos idiomas. Estos factores, junto a la redundancia existente en la Web y las opiniones y hechos contradictorios que aparecen, hacen que los usuarios inviertan mucho más tiempo de lo deseado navegando, buscando y seleccionando la información que es de su interés.

En este sentido, las Tecnologías del Len-

[1] http://www.internetworldstats.com/stats.htm

ISSN 1135-5948

guaje Humano (TLH) son clave para facilitar al usuario la gestión de toda esta información. Actualmente la investigación en esta área suele centrarse en una tarea específica e independiente, como puede ser la recuperación de información (Vila et al., 2013), minería de opiniones (Fernández, Gómez, y Martínez-Barco, 2010), desambiguación del sentido de las palabras (Gutiérrez et al., 2013) o generación de resúmenes (Vodolazova et al., 2013). Sin embargo, dadas las necesidades del contexto actual, donde la información crece a un ritmo exponencial, es necesario aunar esfuerzos en las distintas tareas hacia la creación de un marco flexible capaz de identificar el tipo de información que necesita el usuario, buscarla, procesarla y presentársela de manera adecuada, para que, por un lado, le ahorre tiempo de proceso y, por otro, le sea útil respecto a sus intereses.

El objetivo principal de este proyecto de investigación[2] es analizar, proponer y evaluar diferentes enfoques novedosos para la anotación y generación de textos adaptados al usuario, creando un marco inteligente que combine e integre distintas aplicaciones de TLH y sea de referencia para la comunidad investigadora. La generación de textos que se propone en este proyecto es flexible, puesto que el resultado no va a ser siempre un texto con el mismo formato, sino que se obtendrá un paquete de información que contendrá anotaciones a distintos niveles, y permitirá utilizar y extraer aquellas que se consideren más apropiadas según el contexto y las necesidades de los usuarios, como son resúmenes, tuits, valoraciones de opiniones, pasajes, recopilación de fuentes relevantes, etc. Esto permitirá una mejor gestión de la información disponible en Internet, proporcionando al usuario información con mayor o menor detalle sobre los temas que le interesen, que le ayudarán en multitud de tareas, incluyendo la consulta de información y/o novedades, y toma de decisiones.

Un valor añadido del proyecto es que los resultados del marco de TLH podrán ser consumidos tanto por seres humanos como por agentes informáticos. La posibilidad de compatibilizar la salida con agentes informáticos, extiende el umbral de éxitos de la propuesta de marco de TLH hacia los horizontes del mercado industrial y/o empresarial, pues posibilita que se establezcan intereses comunes entre ambas comunidades, la científica y la empresarial.

[2]http://gplsi.dlsi.ua.es/gplsi11/en/node/16595

## 2 *Arquitectura general para el marco de TLH*

En la figura 1 se ilustra el marco TLH propuesto, donde podemos observar cómo este marco permite a los usuarios consultar información de Internet (medios sociales, foros, noticias, etc.) y dependiendo de las necesidades que tenga el usuario, presentarle la información de una u otra manera (por ejemplo, mediante un tuit, un resumen, una valoración de un tema, etc.).

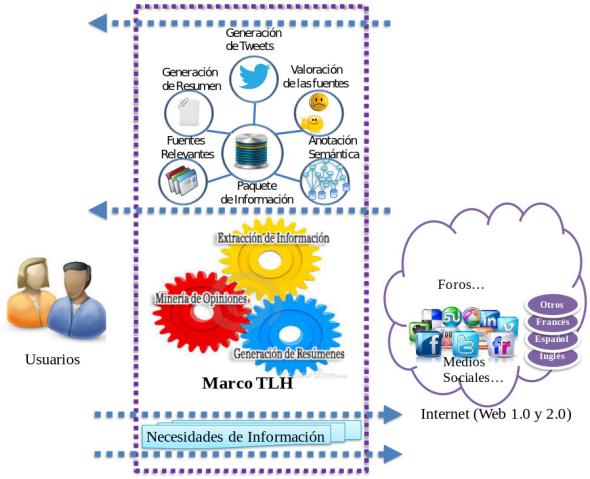


Figura 1: Ilustración del marco de TLH

Tal y como se aprecia en la figura, las herramientas y procesos de TLH son el elemento central y juegan un papel clave desde dos puntos de vista: por un lado, se utilizan para buscar y recuperar la información necesaria de la Web, y por otro, son esenciales en el procesamiento y tratamiento inteligente de dicha información para obtener conocimiento. Concretamente, se integrarán tecnologías como el análisis semántico, la recuperación y la extracción de información, la minería de opiniones y la generación de resúmenes. Aunque el marco no está limitado a la integración de estas aplicaciones, sí que es cierto que estas tareas serán las que conformen su núcleo central, y por tanto, serán cruciales para el correcto desarrollo del proyecto.

Para llevar a cabo la generación de textos debemos en primer lugar decidir qué información recuperar y seleccionarla. Posteriormente se procesará dicha información, ya sea subjetiva u objetiva, y para ello, será necesario

identificar el tipo de información, clasificarla, detectar lo realmente importante, determinar información redundante, complementaria y/o contradictoria e integrar y combinar todo el conocimiento obtenido. Todo este conocimiento obtenido quedará anotado de forma automática en un paquete de información (lo que denominaremos en nuestro proyecto "paquete semántico"), en base a una ontología previamente diseñada. Una vez poblada la ontología, la última fase sería generar un texto que cubra las necesidades de los usuarios y que pueda ser mostrado en base a diferentes formatos, a partir de las anotaciones que contenga. O bien, como se ha comentado en la sección 1 que directamente el documento anotado semánticamente pueda ser procesado y utilizado por otras aplicaciones informáticas.

## *3 Estado del proyecto*

Durante el primer año, nos hemos centrado en la definición y creación del paquete semántico. Este proceso abarca desde el diseño de la ontología para representar distintos tipos de información que puede contener un paquete semántico, hasta el desarrollo de un proceso automático que sea capaz de integrar las herramientas de TLH a utilizar y poblar la ontología de forma automática.

### 3.1 Definición del paquete semántico

Como base para la definición del paquete semántico, hemos diseñado una ontología utilizando la herramienta Protégé[3]. Dicha ontología contiene tanto información léxica como semántica de cómo hemos decidido que se representen los documentos, las frases y el resumen derivado. Junto con la definición y diseño de la ontología, tenemos asociadas un conjunto de preguntas de competencia que serán las que la ontología deberá resolver (por ejemplo, *"¿qué resúmenes refieren hechos que datan del día dd/mm/aaaa?"*o *"¿qué entidades nombradas están implicadas en documentos del dominio deportivo?"*).

La figura 2 muestra la jerarquía de clases definidas en el paquete semántico.

### 3.2 Creación del paquete semántico

Una vez definida y diseñada la ontología, la fase de creación del paquete semántico está compuesta por tres módulos que se ejecutarán de manera secuencial, y que juntos van a constituir el núcleo central del marco de TLH. A continuación, se explicará cada uno de estos módulos con más detalle.

#### 3.2.1 Gestor de fuentes de información

En primera instancia, necesitamos disponer del conjunto de fuentes de información de partida. Por lo tanto, este primer módulo tiene como objetivo descargar las fuentes de información con las que se desee trabajar y extraer el texto que posteriormente se procesará.

#### 3.2.2 Integrador de procesos de TLH

Este módulo es el encargado de ejecutar los procesos de TLH deseados y determinar las entradas y salidas de cada uno. Para un primer prototipo, hemos seleccionado un conjunto de herramientas de TLH para poder procesar los documentos. Como premisa, hemos optado en la medida de lo posible reutilizar herramientas ya existentes en cada una de las áreas que han demostrado ser competitivas en su ámbito. Estas herramientas se resumen en la tabla 1 y todas ellas funcionan para el idioma inglés[4].

| Proceso TLH | Herramienta | Institución |
|---|---|---|
| A.Semántico | ISR-Wordnet | UA |
| A.Sentimientos | Sentiment | UA |
| Gen.Resúm | GPLSICompendium | UA |
| Rec.NER | StandfordNER | Standford |
| Rec.ExprTemp | TipSem | UA |

Tabla 1: Procesos de TLH integrados

#### 3.2.3 Anotación semántica

La finalidad de este módulo es poblar la ontología previamente diseñada en base a la información proporcionada por los procesos de TLH. Como resultado de ejecutar estos tres módulos, tendremos ya creado el paquete semántico listo para poder hacer diferentes consultas en función de las necesidades de información o bien para ser integrado en otros procesos automáticos. A modo ilustrativo, un ejemplo de algunos componentes que integrarían el paquete semántico pueden verse gráficamente en la figura 3.

[3] http://protege.stanford.edu/

[4] Algunas de estas herramientas están accesibles a través de: http://gplsi.dlsi.ua.es/services/pln/doc/index.html

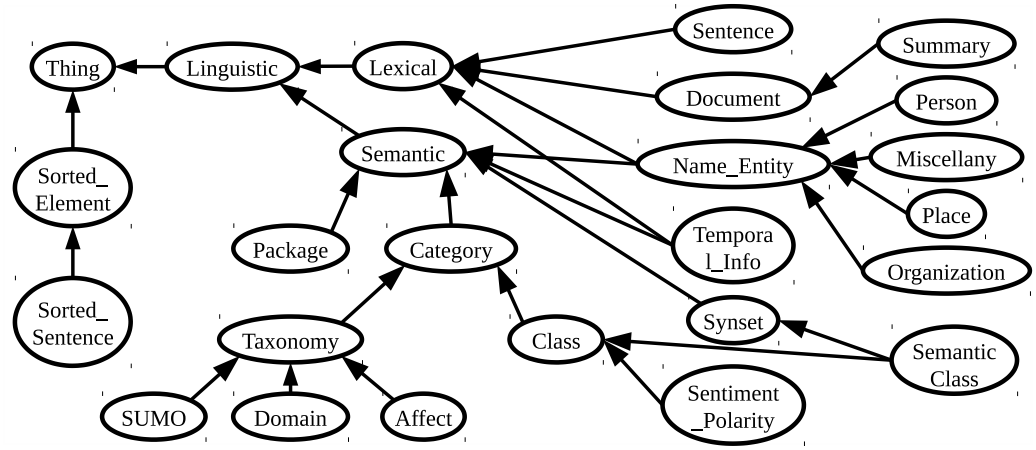


Figura 2: Jerarquía de clases de la ontología para un paquete semántico

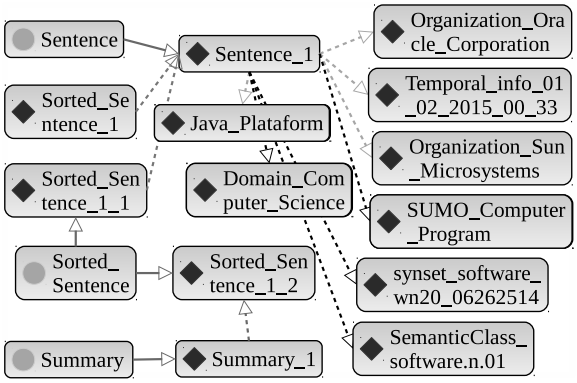


Figura 3: Ejemplo de una frase y los conceptos léxico-semánticos que contiene

## *4 Trabajo futuro*

De cara al segundo año, nos vamos a centrar en analizar y estudiar métodos para presentar la información del paquete semántico de manera flexible y adaptada, para evaluarla mediante estudios de usuario. También estudiaremos la posibilidad de integrar y aplicar directamente el paquete semántico a otras posibles tareas de TLH. A la finalización del proyecto se pretende dejar accesible el marco de TLH desarrollado a través de la API de servicios del grupo GPLSI, así como la ontología diseñada para la creación del paquete semántico.

## *Agradecimientos*

Este proyecto ha sido financiado por la Universidad de Alicante a través del proyecto emergente "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15) y su temática se enmarca en el contexto de los proyectos "DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y Generación de Información sobre la Web 2.0" (PROMETEOII/2014/001) financiado por la Generalitat Valenciana y el proyecto "Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano" (TIN2012-31224) financiado por Ministerio de Economía y Competitividad del Gobierno de España.

## *Bibliografía*

Fernández, J., J. M. Gómez, y P. Martínez-Barco. 2010. Evaluación de sistemas de recuperación de información web sobre dominios restringidos. *Procesamiento del lenguaje natural*, 45:273–276.

Gutiérrez, Y., Y. Castaneda, A. González, R. Estrada, D. D Piug, J. I. Abreu, R. Pérez, A. Fernández Orqun, A. Montoyo, R. Muñoz, y F. Camara. 2013. UMCC DLSI: Reinforcing a ranking algorithm with sense frequencies and multidimensional semantic resources to solve multilingual word sense disambiguation. *Proc. of SemEval*, páginas 241–249.

Vila, K., A. Fernández, J. M. Gómez, A. Ferrández, y J. Díaz. 2013. Noise-tolerance feasibility for restricted-domain information retrieval systems. *Data & Knowledge Engineering*, 86:276–294.

Vodolazova, T., E. Lloret, R. Muñoz, y M. Palomar. 2013. Extractive text summarization: Can we use the same techniques for any text? En *Natural Language Processing and Information Systems*. Springer, páginas 164–175.