risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 18/1/2016
Aceitação/Acceptance: 19/5/2016

TLH Suite: herramienta para la anotación semántica de información

Antonio Guillén, Elena Lloret, Yoan Gutiérrez

{aguillen, elloret, ygutierrez}@dlsi.ua.es

Grupo de Procesamiento del Lenguaje y Sistemas de Información, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Carretera San Vicente del Raspeig S/N, 03690, San Vicente del Raspeig, España.

DOI: 10.17013/risti.18.99–113

Resumen: En la actualidad existe gran cantidad de información heterogénea en Internet, esto dificulta que los usuarios puedan encontrar y filtrar fácilmente la información que requieren. Las herramientas basadas en las Tecnologías del Lenguaje Humano (TLH) ofrecen un gran apoyo facilitando la tarea y proporcionando al usuario la información específica que solicita. El objetivo de este artículo es proponer una herramienta capaz de procesar y anotar la información textual proveniente de la Web. Dicha herramienta viene motivada precisamente por la necesidad de definir un marco tecnológico que consiga integrar una serie de recursos TLH existentes, de manera que se genere un paquete de información semántica que pueda ser consultado flexibilizando el tipo de información a recopilar dadas las necesidades del usuario. Los resultados obtenidos en la experimentación demuestran el valor añadido que aporta el marco propuesto con respecto al uso de los recursos TLH de forma individual.

Palabras-clave: Tecnologías del lenguaje humano, Paquete semántico, Generación de textos, Anotación, Internet

TLH Suite: semantic information annotation tool

Abstract: Nowadays, the vast amount of heterogeneous information available on the Internet poses difficulties for users when they have to find the information they require, since this is a non-trivial task. In this respect, Human Language Technologies (HLT) tools offer a great support for this task, being able to provide the specific information requested by the user. The aim of this paper is to propose a tool capable of processing and annotating the textual information from the Web. This tool is motivated by the need of defining a technological framework to integrate a number of existing HLT resources, so that a semantic information package is generated. This package may also adapt the type of information to retrieve and generate, based on the particular user needs. The results obtained from the experimentation performed show the added value brought by our proposed HLT framework compared to the use of individual HLT resources.

Keywords*:* Human language technologies, Semantic package, Text generation, Annotation, Internet

1. Introducción

La llamada *Sociedad de la Información* no sólo exige el acceso a todo tipo de materiales sino que en la actualidad genera gran cantidad de contenido que toma especial relevancia en actividades sociales, culturales y económicas. Por lo tanto se hace cada vez más complicado el procesamiento y tratamiento de la información, especialmente si el usuario desea centrarse con mayor o menor detalle en un tema concreto. Dicha información se encuentra en fuentes 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 Internet existen alrededor de 3.300 millones de usuarios conectados lo que supone más del 40 % de la población mundial[1]. Desde la aparición de la Web 2.0 (o Web social), se han creado nuevos sitios Web, como por ejemplo las redes sociales, foros, o microblogs, donde los usuarios juegan un papel más activo pudiendo participar, interactuar e intercambiar información con otros usuarios.

En este sentido, las *Tecnologías del Lenguaje Humano* (TLH) son necesarias para ayudar al usuario a gestionar la gran cantidad de información que ofrece Internet. Las herramientas TLH pueden llegar a revolucionar la manera de procesar la información y conseguir establecer procesos más eficientes y eficaces. Actualmente la investigación en esta área suele centrarse en tareas específicas e independientes, como puede ser: la recuperación de información (Vila et al., 2013), generación de resúmenes (Vodolazova et al., 2013), desambiguación del sentido de las palabras (Gutiérrez et al., 2013) o la minería de opiniones (Fernández et al., 2010). El objetivo de este trabajo es presentar, describir y evaluar TLH Suite, una herramienta que permita integrar una serie de recursos TLH existentes. Por una parte anota automáticamente un documento utilizando los recursos TLH integrados. Por otra parte almacena toda la información anotada de forma estructurada generando un paquete de información semántica que pueda ser consultado de forma flexible y selectiva por parte del usuario.

Más allá del objetivo básico, pretendemos tener una herramienta capaz aplicarse a cualquier tipo de documento digital que consideren los usuarios. Estos documentos pueden ser perfectamente contenidos de la Web, de los cuales se podría realizar una anotación semántica. El análisis de estos datos semánticos facilitará la comprobación de la información tratada en redes sociales, foros, blogs, Webs de comercio electrónico, etc. Esta tarea es de especial relevancia principalmente para asegurar que los sitios Web respetan los derechos de los usuarios y la legalidad (Jiménez et al., 2013). También puede ser útil dentro de la administraciones públicas u otras instituciones, por el motivo de que cada vez más usuarios hacen uso de las *Tecnologías de la Información y la Comunicación* (TIC) para el acceso a los servicios sociales o de salud pública (Cruz-Cunha et al., 2014). En este aspecto, el uso de nuestra herramienta podría proporcionar una mejora en la gestión de estos servicios derivando en una atención a los destinatarios más rápida, flexible y efectiva.

[1] Datos del 30 de noviembre de 2015 provenientes de: http://www.internetworldstats.com/stats.htm (consultado el 14 de enero de 2016)

La estructura del artículo es la siguiente: en la sección 2 se revisa la literatura y las contribuciones que se han hecho a esta investigación; la sección 3 presenta el diseño de la herramienta describiendo detalladamente sus características y funcionalidades; en la sección 4 se expone la experimentación y evaluación que se ha llevado a cabo así como la discusión de los resultados; en la sección 5 se exponen las conclusiones y el trabajo futuro.

2. Revisión de literatura y contribuciones de esta investigación

Las *Tecnologías del Lenguaje Humano* consisten principalmente en gestionar de forma automática cualquier fuente escrita o hablada generalmente no estructurada, como textos y conversaciones siendo de cualquier origen y dominio de aplicación. Dado que el propio lenguaje humano es complejo a nivel estructural y cognitivo, no es un área de investigación trivial. Un recurso TLH debe usar abundante conocimiento de las estructuras del lenguaje humano pero también sobre el significado de las palabras u oración completa, así como conocimiento del mundo para poder obtener la información verdaderamente relevante y fiable. Debido a que es un área de investigación que representa un gran reto, necesariamente se ha de fraccionar en tareas más o menos complejas. Además, ante la gran cantidad de información existente en Internet, esta área cobra mayor relevancia, y sus recursos derivados de tareas comunes son altamente solicitados.

Algunas de las tareas más importantes de TLH relacionadas con Internet son: Recuperación y extracción de información (Vila et al., 2013), Análisis semántico mediante el recurso ISR-Wornet (Gutiérrez et al., 2010), Generación de resúmenes mediante el recurso Compendium (Lloret & Palomar, 2012), Minería de opiniones o análisis de sentimientos principalmente en redes sociales (Fernández et al., 2015), Reconocimiento de expresiones temporales con el recurso TIPSem (Llorens et al., 2010). Otras tareas que hoy en día empiezan a ser relevantes principalmente para Internet y el auge de la *Web Semántica*[2] son: la obtención de características basadas en análisis semántico (Dávila et al., 2012) principalmente para la generación de meta-datos en la *Web Semántica*, la detección del grado de dificultad de lectura (Martín-Valdivia et al., 2014) que es de interés para la clasificación de contenidos, la identificación de autoría (Sapkota et al., 2015) muy útil para la generación de perfiles de usuarios, detección de género y edad (Rangel & Rosso, 2016) también útil para la generación de perfiles de usuarios y la clasificación de contenidos.

La idea de este trabajo es realizar una herramienta que integre varios de los recursos mencionados en el párrafo anterior. Como antecedente a esta herramienta tenemos InTiMe (Gómez, 2008) un proyecto del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante. Esta herramienta trata de integrar gran cantidad de recursos TLH. De este modo muchos investigadores tendrán acceso a todos de forma remota e independientemente del sistema operativo. Así mismo, permite a desarrolladores e investigadores incorporar sus propios recursos a la herramienta de forma sencilla. Desgraciadamente, esta herramienta solo permite el uso individual de

[2] La Web Semántica se ocupa de vincular datos a la World Wide Web que sean legibles por aplicaciones informáticas.

los recursos sin existir conexión alguna entre ellos. También existe otra herramienta web llamada MeaningCloud[3] que aparte de integrar varios recursos TLH, establece conexiones entre la información anotada por todos ellos.

Como se ha explicado, las herramientas expuestas anteriormente contemplan limitaciones o no responden a las necesidades y expectativas que se pretenden cumplir con la propuesta. En el caso de la herramienta InTiMe aunque integra varios recursos solo se puede procesar individualmente con uno a la vez. Con respecto a MeaningCloud aunque si cumple el requisito de integrar varios recursos y poder procesar con todos ellos a la vez, a la hora de recuperar información no es posible recopilarla forma selectiva y metódica por ejemplo mediante consultas. Por lo tanto estas herramientas no son capaces de proporcionar toda la potencia y flexibilidad que requiere el usuario. Con la propuesta de este artículo damos un paso más allá, permitirá integrar y conectar diferentes recursos TLH y será capaz de proporcionar la información semántica específica que solicite el usuario mediante consultas. Esto se conseguirá realizando conexiones entre las anotaciones provenientes de los recursos mediante el paquete semántico para poder facilitar las consultas que pueda realizar el usuario.

Otro valor añadido con respecto a las soluciones existentes, es que utilizaremos recursos TLH que se basan de métodos y técnicas ampliamente utilizadas en el mundo científico, lo que garantiza una estabilidad y calidad en el funcionamiento al estar avalados por publicaciones o tesis. Además permitiremos que sea fácilmente adaptable a nuevas versiones y mejoras de los recursos existentes utilizando nuevos métodos, otorgando a la herramienta un carácter evolutivo. Asimismo también pretendemos que la herramienta sirva de apoyo a los sistemas de recomendación (Bobadilla et al., 2013) que se usan principalmente para la recomendación de productos a los usuarios del comercio electrónico. El papel que tomaría nuestra herramienta en este ámbito, es la ayuda a la generación automática de perfiles de usuario a partir de la anotación semántica de los contenidos e información que aporta el mismo en redes sociales y demás sitios Web.

3. Diseño de la herramienta

En esta sección se describirá la herramienta junto a todas las características y funcionalidades añadidas. Se empezará por la arquitectura diseñada, siguiendo con la gestión de las principales funcionalidades de la herramienta.

3.1. Arquitectura

La arquitectura de TLH Suite mostrada en la figura 1 trata de describir la composición y funcionamiento de los módulos, así como la interacción de estos con los elementos de entrada, salida, procesado y almacenamiento. Los módulos que componen la herramienta son los encargados de realizar las funcionalidades principales de la misma. Están desarrollados en Java y dispone cada uno de un ejecutable BASH Script para su uso mediante línea de comandos. La salida de estos se muestra por consola o se guarda

[3] http://www.meaningcloud.com (consultado el 13 de enero de 2016)

en una base de datos MongoBD[4]. El resultado almacenado es lo que se conoce como paquete semántico, esto quiere decir que el texto se queda anotado semánticamente según los recursos que se han utilizado, lo que permite poder filtrar y extraer la información deseada mediante consultas.

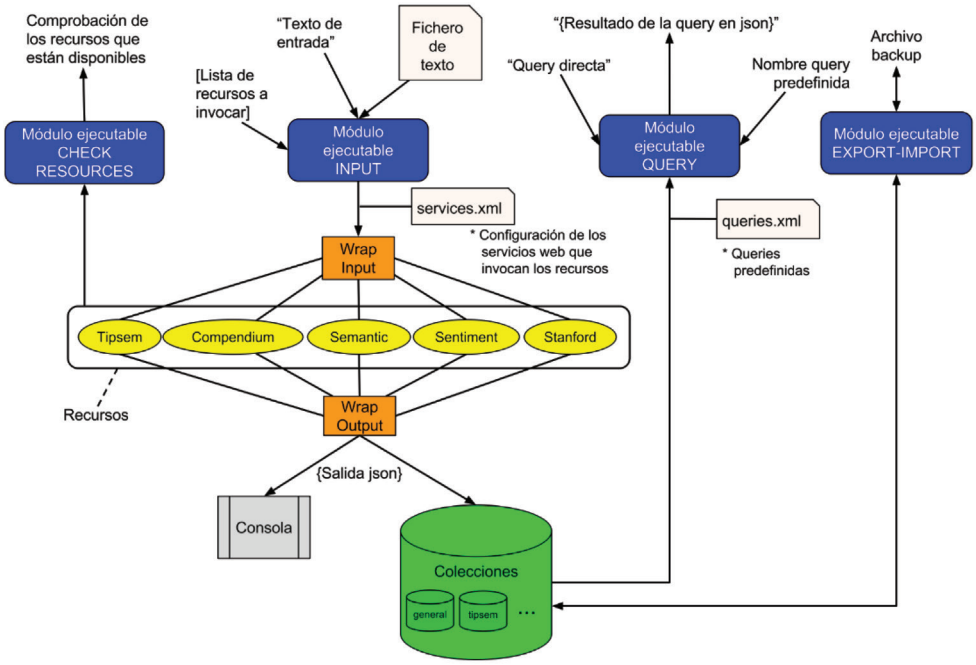

Figura 1 – Arquitectura de TLH Suite

El módulo más importante de la herramienta es el *Módulo Input*. Este se encarga de aceptar la entrada de documentos (ya sean documentos en texto plano o cadenas de texto). A su misma vez se indica la lista de recursos con los que queremos procesar cada documento de la entrada, en esta lista deben aparecer únicamente los recursos definidos en la herramienta mostrados en amarillo, aunque se pueden definir nuevos recursos modificando el fichero donde se definen los servicios web[5] de cada recurso (*services.xml*). Una vez aceptada la entrada, el módulo comienza a procesar cada documento con cada recurso indicado, a su misma vez se va proporcionando la salida que consiste en la anotaciones semánticas realizadas sobre los documentos utilizando los recursos indicados. La salida (paquete semántico) se puede emitir por consola o ser almacenada en la base de datos MongoBD, en este último caso es el *Módulo Query* se encargará posteriormente de realizar las consultas a petición del usuario sobre el paquete semántico almacenado.

[4] https://www.mongodb.org (consultado el 15 de enero de 2016)

[5] Un servicio web es una tecnología que utiliza un conjunto de protocolos y estándares que sirve para usar un recurso informático de forma remota.

El usuario puede realizar estas consultas de dos modos, mediante una consulta directa a la herramienta o invocando una consulta predefinida del fichero *queries.xml* (que el mismo usuario puede personalizar). Los restantes módulos tienen funcionalidades más simples, el *Módulo Import/Export* se encarga de la importación y exportación de paquetes semánticos almacenados, y *Módulo Check Services* se encarga de comprobar el correcto funcionamiento y la accesibilidad a los servicios web de los recursos TLH.

Para llevar a cabo las funcionalidades explicadas y poder incorporar nuevas en el futuro con facilidad, se tienen en cuenta una serie de gestiones principales que intervienen en el funcionamiento completo de la herramienta. Estas gestiones son tres: gestión de recursos, gestión de almacenamiento y gestión de consultas, que se explican más en detalle en las siguientes secciones.

3.2. Gestión de recursos

La gestión de recursos consiste en gestionar la integración y uso de los recursos TLH seleccionados para la herramienta. Estos recursos están encapsulados en servicios web para su uso remotamente. Uno de los criterios que se ha seguido para la selección de recursos es que sean capaces de anotar la información de un documento en base a distintos niveles de detalle. Debido a que algunos recursos solo funcionan para el idioma inglés, se ha decidido que la herramienta funcione solo para este idioma, aunque algunos de los recursos también funcionan para español. En la tabla 1 se muestran los recursos seleccionados e información sobre ellos.

Recurso	Referencia	Tarea TLH	Tipo	Idioma	Público
Tipsem	(Llorens et al., 2010)	Reconocedor de expresiones temporales	REST	Inglés	Sí
Compendium	(Lloret & Palomar, 2012)	Generación de resúmenes	SOAP	Inglés	Sí
Semantic	(Gutiérrez et al., 2010)	Análisis semántico de textos	SOAP	Inglés	No
Sentiment	(Fernández et al., 2015)	Análisis de sentimientos	REST	Inglés Español	Sí
Stanford	(Finkel et al., 2005)	Reconocedor de entidades nombradas	REST	Inglés Español	Sí

Tabla 1 – Selección de recursos para TLH Suite

Los recursos TLH seleccionados son incluidos en la herramienta mediante un fichero de configuración con formato XML que entre otros datos indica la dirección del servicio web asociado para cada recurso. Es necesario implementar un componente llamado *wrapper* que se encargue de invocar el servicio web correspondiente al recurso adaptando sus entradas y sus salidas. La adaptación de las entradas es porque cada servicio web necesita un tipo de entrada distinta. La adaptación de la salida se realiza para poder almacenar

adecuadamente las anotaciones generando correctamente el paquete semántico. La gestión de recursos se realiza en los módulos *Input* y *Check Services*.

3.3. Gestión de almacenamiento

La gestión de almacenamiento trata de definir cómo se genera el paquete semántico a partir de las anotaciones obtenidas por los recursos y cómo se almacena el mismo. Para el almacenamiento del paquete semántico se ha decidido usar una base de datos NoSQL (MongoDB en concreto). Este tipo de base de datos permitirá la adecuada interconexión de información a diferentes niveles de análisis lingüístico para proceder a una mejor y más eficiente consulta de datos. Por lo tanto se ha diseñado una base de datos MongoDB con una serie de colecciones (equivalente a las tablas en una base de datos relacional). En cada colección se guarda en formato JSON[6] las anotaciones resultantes del recurso adaptadas mediante su correspondiente *wrapper*.

En total la herramienta dispone de 7 colecciones de datos MongoDB para almacenar las anotaciones y establecer las conexiones del paquete semántico. Todas las colecciones tienen 2 campos comunes que corresponden al identificador de documento y el identificador de escenario. Las colecciones guardan la entrada de texto original y las anotaciones provenientes de la invocación de cada recurso. Una de las colecciones llamada *general* guarda todas las anotaciones correspondientes a todos los recursos invocados incluyendo el texto original, lo que permite que las consultas de datos sean más fáciles de diseñar utilizando esta colección.

3.4. Gestión de consultas

La gestión de consultas trata de definir una forma intuitiva y amigable de realizar consultas sobre el paquete semántico generado a partir de las anotaciones. Se han creado dos tipos de consultas: nativas y predefinidas. Las consultas nativas se invocan directamente con la herramienta mediante el módulo *Query*. Estas consultas propias de la herramienta tienen ciertas equivalencias con las consultas de bases de datos relacionales pero en formato JSON. Las consultas predefinidas se indican en un fichero XML de configuración para poder ser invocadas mediante un identificador definido para cada una en el mismo fichero.

4. Experimentación y evaluación

El objetivo de la experimentación y evaluación que se va a hacer de la herramienta es comprobar que responde adecuadamente ante diferentes tipos de entrada y uso de recursos, y la salida mediante consultas, verificando la mejora con respecto a las herramientas expuestas como trabajo previo. Para la experimentación se propone anotar una serie de documentos de diversas temáticas, para posteriormente consultar información mediante un conjunto de 50 preguntas de competencia que se lanzarán a la herramienta convertidas en consultas predefinidas. Estos documentos están agrupados en 3 escenarios de prueba.

[6] JSON (JavaScript Object Notation) es un formato simple, ligero y generalizado para el intercambio de datos.

4.1. Escenarios de prueba

Un escenario de prueba es una experimentación basada en un conjunto de documentos de determinada temática, de manera que la herramienta se pueda evaluar mediante una serie de preguntas de competencia orientadas a dicha temática. Para la evaluación se han definido 3 escenarios de prueba distintos, los cuales se resumen en la tabla 2.

Escenario	Temática	Descripción	Documentos	Nº Frases	Nº Palabras
Escenario 1	Gastronomía	Artículo sobre la apertura de McDonalds en la Unión Soviética	1 texto largo	47	1052
Escenario 2	Deportes	Noticia sobre un partido de tenis de Rafa Nadal	2 textos cortos	15	281
				doc1: 9	**doc1**: 157
				doc2: 6	**doc2**: 124
Escenario 3	Cine	Biografía del director de cine Stanley Kubrick	2 textos largos	119	2053
				doc1: 52	**doc1**: 1115
				doc2: 67	**doc2**: 938

Tabla 2 – Escenarios de prueba

El hecho de crear escenarios de distinta temática da la posibilidad también de evaluar el comportamiento de la herramienta ante diversas áreas de interés. Así mismo el tamaño y la cantidad de documentos seleccionados para cada escenario variarán para comprobar el rendimiento de la herramienta. Los documentos creados para los escenarios han sido extraídos de páginas web de noticias y cine[7]. Cada documento es un fichero separado de texto plano en idioma inglés. Un fragmento de los documentos de cada escenario se puede ver en la figura 2.

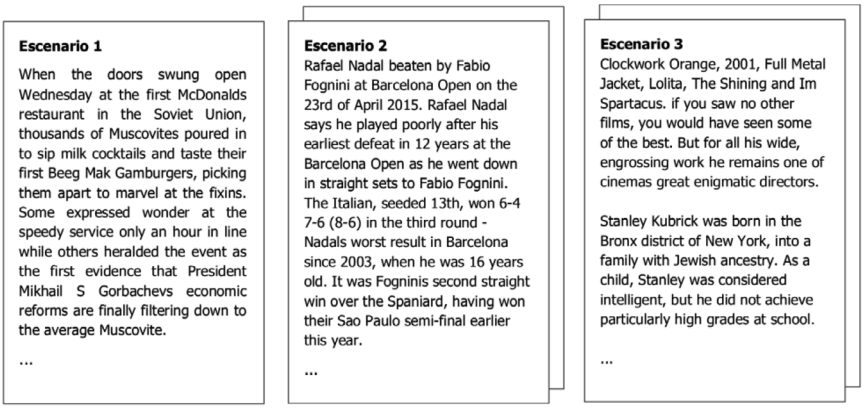

Figura 2 – Contenido resumido de los escenarios

[7] El contenido completo de los documentos de los escenarios se puede consultar en: https://docs.google.com/document/d/16jq-U1EFrKKheXRLQkdl8OpJGPP-eo-ppUG4EbaDVpY/edit?usp=sharing

La ejecución de la herramienta se hace por cada uno de los documentos. El procesado que lleva a cabo la herramienta es el siguiente: primero dividir el documento entrante en sentencias separadas (entre puntos) y cada recurso TLH será invocado a nivel de sentencia exceptuando los recursos que trabajan a nivel de documento (solo Compendium). Para esta experimentación se utilizarán todos los recursos TLH disponibles en la herramienta, para poder hacer consultas combinando criterios de más de un recurso.

La experimentación se ha llevado a cabo de esta manera: primero se lanza el módulo Input con cada uno de los documentos de los 3 escenarios usando los 7 recursos TLH disponibles. Una vez generado el paquete semántico correspondiente a los 3 escenarios se procederá a transformar las preguntas de competencia en consultas predefinidas. Finalmente se lanzará el conjunto de 50 consultas predefinidas con el módulo *Query* para comprobar los resultados obtenidos.

4.2. Preguntas de competencia

Las preguntas de competencia son preguntas textuales en lenguaje natural. Se definen en base a las necesidades de información más comunes que un usuario quiere extraer con la herramienta. En el caso concreto de la evaluación se ha pensado y preparado un conjunto 50 preguntas de competencia que contemple diversos aspectos a evaluar de la herramienta: su capacidad para abordar diferentes dominios y temáticas, abarcar diversos criterios de filtrado de información o niveles de análisis lingüístico, etc. Por lo tanto se puede dividir el conjunto en 3 tipos de preguntas[8]:

- Preguntas simples (9 consultas).
- Preguntas combinadas (37 consultas).
- Preguntas complejas (4 consultas).

Las consultas simples consisten en preguntar un criterio que corresponda solo a un recurso TLH, como por ejemplo una pregunta donde solo se usa un criterio del recurso de expresiones temporales Tipsem:

"Expresiones temporales que aparecen"

Las consultas combinadas abarcan criterios de 2 o más recursos en la pregunta. Un ejemplo de pregunta combinada sería la que combina criterios de los recursos Sentiment (para obtener frases positivas) y Stanford (para obtener frases que contienen la entidad número):

"Frases positivas que contengan alguna entidad NUMBER"

Las consultas complejas ya tratan un cierto nivel de análisis lingüístico y una alta conexión de información, como por ejemplo la conexión entre expresiones temporales y las entidades persona:

"Fechas asociadas a una persona"

[8] El listado completo con las preguntas definidas se puede ver en: https://docs.google.com/document/d/1BkLTCtiDFpKq4MezgSKMGR74NK40YgiV-Lv69qXS3e0/edit?usp=sharing

Finalmente, comentar que las preguntas de competencia no pueden ser directamente lanzadas a la herramienta porque están escritas en lenguaje natural. Previamente se deben transformar a un formato de consulta predefinido de TLH Suite[9].

4.3. Resultados

Se ha realizado la evaluación de la herramienta usando el módulo *Query* con las 50 consultas de TLH Suite generadas a partir de las preguntas de competencia. Todos los resultados han sido guardados y verificados manualmente para comprobar el número de consultas que puede resolver la herramienta[10]. Así mismo, se ha hecho una experimentación paralela utilizando los recursos de forma individual sin utilizar TLH Suite, comprobando también en qué casos se puede resolver la consulta. Todos estos datos se resumen en la tabla 3.

Tipo de consulta	Cantidad	Resueltas con TLH Suite	Resueltas sin TLH Suite
Simples	9	9	9
Combinadas	37	37	0
Complejas	4	0	0
Total	50	46	9

Tabla 3 – Resultados por tipo de consulta

En total 46 de las 50 consultas se pueden resolver usando TLH Suite y solo 9 se pueden resolver utilizando individualmente los recursos TLH (sin TLH Suite). Ambas formas permiten resolver las consultas simples, pero solo con TLH Suite se pueden resolver las consultas combinadas. A continuación se va a analizar algunos de los resultados de los 3 tipos de consultas.

En cuanto a las consultas simples, tal y como se puede ver en el ejemplo de la figura 3, solo se utiliza un criterio de un recurso TLH que en este caso es el de expresiones temporales Tipsem. El resultado emite todas las expresiones temporales, cada una de estas con su término textual original y su valor correspondiente anotado.

En la tabla 4 se puede observar una muestra de 4 consultas simples de la evaluación.

#	Consulta	Con TLH Suite	Sin TLH Suite
01	Entidades lugar	✓	✓
02	Entidades persona	✓	✓
03	Expresiones temporales que aparecen	✓	✓
04	Número de frases positivas	✓	✓

Tabla 4 – Muestra de 4 consultas simples

[9] Las consultas TLH Suite generadas a partir de las preguntas se puede ver en: https://drive.google.com/file/d/0B1E7Lgk0ANquR240S1Rfc2NfYnM/view?usp=sharing

[10] Los resultados de las 50 consultas realizadas con TLH Suite se pueden encontrar en: https://docs.google.com/document/d/1iL7x6_gwEvNAQkW9GITtetE0mc1ek-PRpE9hULlnA-g/edit?usp=sharing

"Expresiones temporales que aparecen"

Query predefinida:

```
<collection>general</collection>
<fields>{tipsem:1,_id:0}</fields>
<filter>{tipsem.times:{$ne:null}}</filter>
```

Resultado reducido:

```
{
    "times": [
        { "value": "2015-09-23", "term": "Wednesday" },
        { "term": "14 years", "value": "P14Y" },
        { "term": "nine months", "value": "P9M" },
        { "value": "2015-09-23", "term": "Wednesday" },
        { "term": "recent years", "value": "PAST_REF" },
        ...
```

Figura 3 – Ejemplo de consulta simple

Por otro lado, las consultas combinadas abarcan varios criterios como se ha comentado en la sección anterior. En la figura 4 podemos ver un ejemplo de consulta combinada, en la que se utilizan 2 recursos: el recurso Sentiment (filtrando únicamente las frases positivas) y el recurso Stanford (filtrando únicamente las frases con entidades NUMBER). Por lo tanto, las frases del resultado cumplirán ambos requisitos.

"Frases positivas que contengan alguna entidad NUMBER"

Query predefinida:

```
<collection>general</collection>
<fields>{_id:0,stanford.entities.$:1,sentiment:1,phrase:1}</fields>  <filter>{$and:
[{stanford.entities.NER:'NUMBER'},{sentiment.category:'positive'}]}</filter>
```

Resultado reducido:

```
{
  "phrase": "It tasted very, I would say, unusual, Lubov Sereda, 45, said with a smile.",
  "phrase": "It was very interesting, that gamburger, Zvetlana Generotova, 25, said,
            leaving with a large paper bag filled with takeout food, also a new concept
            here.",
  "phrase": "Fognini levelled again at 55 thanks to a screaming forehand winner off a
            Nadal smash, and then forged into a 63 lead in the tiebreak.",
  ...
```

Figura 4 – Ejemplo de consulta combinada

En la tabla 5 se pueden observar 4 de las consultas combinadas de la evaluación. Evidentemente este tipo de consultas no se pueden resolver utilizando los recursos individualmente, ya que involucran las anotaciones de 2 o más recursos.

#	Consulta	Con TLH Suite	Sin TLH Suite
22	Dominios de frase que contengan alguna referencia temporal	✓	X
23	Frases positivas que contengan alguna entidad NUMBER	✓	X
24	Frases negativas que tengan alguna referencia temporal y la entidad LOCATION	✓	X
25	Frases que contengan la entidad LOCATION y no contengan alguna referencia temporal	✓	X

Tabla 5 – Muestra de 4 consultas combinadas

Finalmente, las consultas complejas tratan criterios de alto nivel lingüístico y conexión de información, por lo tanto no pueden ser resueltas con la versión actual de TLH Suite. En la tabla 6 se puede observar las 4 consultas complejas de la evaluación.

#	Consulta	Con TLH Suite	Sin TLH Suite
47	Todas las expresiones temporales distintas	X	X
48	Entidades nombradas NUMBER coincidentes en el escenario 2 y 3	X	X
49	Fechas asociadas a una persona	X	X
50	Polaridad del escenario completo	X	X

Tabla 6 – Muestra de 4 consultas complejas

4.4. Discusión de los resultados

Con la evaluación que se ha hecho de la herramienta se pretende demostrar que TLH Suite alcanza un nivel más de refinamiento a la hora de consultar la información almacenada. Como se puede observar, TLH Suite puede resolver consultas simples y combinadas debido a que el paquete semántico realiza varias conexiones entre las anotaciones almacenadas. Usando aproximaciones anteriores solo se conseguía una serie de resultados provenientes de recursos TLH individuales, o en el caso de que se integraran estos recursos conectando la información, no teníamos la suficiente flexibilidad para poder separar la información relevante de los resultados. Este hecho supone que no se pueden resolver consultas combinadas con estas aproximaciones anteriores, a menos que se realice un tratamiento manual para filtrar y seleccionar la información relevante. Así mismo, el formato JSON de salida es simple, manejable e intuitivo, además de ser usado comúnmente por muchas aplicaciones actuales para permitir el intercambio de datos.

Sin embargo, en TLH Suite todavía falta mayor conexión entre datos, mayor nivel de análisis y de funcionalidad para permitir resolver las consultas complejas. La consulta 47 no se puede resolver porque TLH Suite no tiene la funcionalidad para omitir los resultados repetidos (equivalente a DISTINCT de MySQL). En el caso de la 48 no se puede resolver porque las consultas no soportan las combinaciones de subconjuntos

de resultados. La consulta 49 no se puede resolver porque no existe una relación entre una expresión temporal de Tipsem y una entidad PERSON del recurso Stanford, y por tanto no existe una asociación de estos conceptos. La 50 no se puede resolver porque no se evalúa ni se guarda el análisis de sentimientos de todo el texto del escenario. Estos detalles se irán resolviendo a medida que avance el proyecto y sean incorporadas nuevas funcionalidades.

5. Conclusiones y trabajo futuro

TLH Suite cumple con el objetivo de procesar y anotar documentos provenientes de Internet utilizando conjunto de recursos TLH existentes. El paquete semántico generado a partir de las anotaciones proporciona un amplio nivel de análisis lingüístico e interconexión de información. El diseño de consultas predefinidas permite que el usuario pueda recuperar información de forma selectiva y metódica. Por otra parte, la arquitectura y el diseño de la herramienta permiten flexibilizar la incorporación de nuevas de características y da la posibilidad de añadir nuevos recursos TLH, permitiendo así afrontar retos futuros y cubriendo nuevas necesidades de los usuarios.

Se han propuesto una serie de ideas para mejorar la herramienta: incorporar una ontología[11], realizar una interfaz web y mejorar la gestión de recursos. Incorporar una ontología para el almacenamiento de datos mejoraría las conexiones entre todas las anotaciones almacenadas y aportaría un mayor nivel de análisis lingüístico. Realizar una interfaz web daría facilidades al usuario a la hora de utilizar la herramienta y evaluar los resultados que ha obtenido. Una mejora de la gestión de recursos consiste en buscar una nueva forma de incorporar los nuevos recursos TLH sin tener que recurrir a la implementación de *wrappers* como se hace actualmente.

Agradecimientos

Esta investigación ha sido financiada por la Universidad de Alicante mediante el proyecto "*Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario"* (GRE13-15), parcialmente financiada por la Generalitat Valenciana a través del proyecto *"DIIM2.0: Desarrollo de técnicas inteligentes e interactivas de minería y generación de información sobre la web 2.0"* (PROMETEOII/2014/001), por el Gobierno de España (MINECO) a través de los proyectos TIN2015-65100-R, TIN2015-65136-C2-2-R, y por la Comisión Europea a través del proyecto SAM (FP7-611312). También queremos agradecer al programa de Formación de Profesorado Universitario de la Universidad de Alicante (FPU-UA) por su apoyo a través de una de sus becas destinada a la formación predoctoral (UAFPU2015-5999).

[11] Definición formal de tipos, propiedades, y relaciones entre entidades que realmente o fundamentalmente existen para un dominio de discusión en particular.

Referencias

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). *Recommender systems survey*. Knowledge-Based Systems, 46, 109–132. doi:10.1016/j.knosys.2013.03.012

Cruz-Cunha, M. M., Simões, R., Varajão, J., & Miranda, I. (2014). *O impacto da exclusão digital na utilização potencial de um mercado eletrónico de serviços de cuidados de saúde e serviços sociais*. RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação, 14, 33–49. doi:10.17013/risti.14.33–49

Dávila, H., Fernández, A., Gutiérrez, Y., Muñoz, R., Montoyo, A., & Vázquez, S. (2012). *Semantic information extraction method on ontologies*. On Conference SEPLN'12, XXVIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, Spain.

Fernández, J., Gómez, J. M., & Martínez-Barco, P. (2010). *Evaluación de sistemas de recuperación de información web sobre dominios restringidos*. Procesamiento del Lenguaje Natural, 45, 273–276.

Fernández, J., Gutiérrez, Y., Gómez, J. M., & Martínez-Barco, P. (2015). *Social rankings: análisis visual de sentimientos en redes sociales*. Procesamiento del Lenguaje Natural, 55, 199–202.

Gómez, J. M. (2008). *InTiMe: plataforma de integración de recursos de PLN*. Procesamiento del Lenguaje Natural, 40, 83–90.

Gutiérrez, Y., Castaneda, Y., González, A., Estrada, R., Piug, D. D., Abreu, J. I., Pérez, R., Fernández-Orquín, A., Montoyo, A., Muñoz, & R., Camara, F. (2013). *UMCC DLSI: Reinforcing a ranking algorithm with sense frequencies and multidimensional semantic resources to solve multilingual word sense disambiguation*. Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, Georgia, USA.

Gutiérrez, Y., Fernández, A., Montoyo, A., & Vázquez, S. (2010). *Enriching the integration of semantic resources based on WordNet*. Procesamiento del Lenguaje Natural, 47, 249–257.

Finkel, J. R., Grenager, T., & Manning, C. (2005). *Incorporating non-local information into information extraction systems by Gibbs sampling*. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05. doi:10.3115/1219840.1219885

Jiménez, D. L., Redchuk, A., Dittmar, E. C., & Vargas, J. P. (2013). *Los logotipos de privacidad en Internet: percepción del usuario en España*. RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação, 12, 49–63. doi:10.4304/risti.12.49–63

Llorens, H., Saquete, E., & Navarro, B. (2010). *Temporal Expression Identification Based on Semantic Roles*. Lecture Notes in Computer Science, 230–242. doi:10.1007/978-3-642-12550-8_19

Lloret, E., & Palomar, M. (2012). *Compendium: a text summarisation tool for generating summaries of multiple purposes, domains, and genres*. Nat. Lang. Eng., 19(02), 147–186. doi:10.1017/s1351324912000198

Martín-Valdivia, M. T., Martínez-Cámara, E., Barbu, E., L., Ureña-López, L. A., Moreda, P., & Lloret, E. (2014). *Proyecto FIRST: Desarrollo de una herramienta para ayudar a personas con autismo mediante la simplificación de textos*. Procesamiento del Lenguaje Natural, 53, 143–146.

Rangel, F., & Rosso, P. (2016). *On the impact of emotions on author profiling*. Information Processing & Management, 52(1), 73–92. doi:10.1016/j.ipm.2015.06.003

Sapkota, U., Bethard, S., Montes, M., & Solorio, T. (2015). Not All Character N-grams Are Created Equal: A Study in Authorship Attribution. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. doi:10.3115/v1/n15-1010

Vila, K., Fernández, A., Gómez, J. M., Ferrández, A., & Díaz, J. (2013). *Noise-tolerance feasibility for restricted-domain Information Retrieval systems*. Data & Knowledge Engineering, 86, 276–294. doi:10.1016/j.datak.2013.02.002

Vodolazova, T., Lloret, E., Muñoz, R., & Palomar, M. (2013). *Extractive Text Summarization: Can We Use the Same Techniques for Any Text?*. Lecture Notes in Computer Science, 164–175. doi:10.1007/978-3-642-38824-8_14