Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 117-124 recibido 01-12-2016 revisado 15-01-2017 aceptado 09-02-2017

Extracción de información temporal de la DBpedia: propuesta de integración en un corpus semiestructurado

Extraction of temporal information of the DBpedia: Integration proposal in a semi-structured corpus

Adolfo Merás, Ana García Serrano y **Ángel Castellanos**
ETSI Informática, UNED
C/Juan del Rosal 16
28040 Madrid
adolfo@meras.com.es, {agarcia, acastellanos}@lsi.uned.es

Resumen: En este trabajo, se hace una propuesta para la extracción automática de información temporal en la DBpedia, suficientemente general para ser aplicada a diferentes dominios. Se experimenta en un dominio concreto, para el que se identificarán y gestionarán recursos DBpedia relacionados. Con la información temporal extraída de los recursos, se alimentará una línea de tiempo y se intersecará a su vez con la información temporal extraída del dominio, en este caso del corpus DIMH (textos semiestructurados o fichas). A continuación, se enriquecerán las fichas originales con la información temporal y se visualizarán y accederá a los resultados organizados sobre la base de su dimensión léxica y temporal. Ante la ausencia de un *gold standard* para evaluar intrínsecamente la propuesta, se aplican criterios dependientes del dominio y de los usuarios y se pone a disposición de la comunidad científica (*GitHub*) el corpus anotado temporalmente.
Palabras clave: DBpedia, extracción - recuperación y acceso a la información, datos abiertos, análisis de conceptos formales.

Abstract: The goal of this work is to make a proposal for the automatic extraction of temporal information in the DBpedia, general enough to be applied to different domains. The experiment is performed using a concrete domain by the identification and management of domain related DBpedia resources. With the relevant temporal information extracted from the resources it will be feed a timeline and intersected with the temporal information of the DIMH corpus (semi-structured texts or cards). Thus, we will enrich these cards with related events of the timeline. In order to visualize the results, we are using a graphical interface to facilitate the lexical and the temporal information access. In the absence of a gold standard to intrinsically evaluate the proposal, it will be applied domain and users dependent criteria and the annotated corpus is made available to the scientific community (GitHub).
Keywords: DBpedia, extraction – retrieval and access to information, open data, formal concept analysis.

1 Introducción

Comprender la información de un dominio concreto exige la construcción de un contexto, en el que la dimensión temporal debe hacerse explícita (Tran et al., 2015), tanto sobre entidades (personajes, obras etc.) como sobre conceptos históricos. En un sistema de acceso a la información, aunque la información temporal "perfecta" la aportaría un experto humanista, la contextualización temporal automática de un corpus puede aprovechar recursos de la Web de los Datos, como es la DBpedia (Zang et al., 2015). En este trabajo se propone contextualizar y enriquecer la información de un corpus de textos con información temporal extraída automáticamente de la DBpedia, mediante:

1. La identificación de los recursos de la DBpedia relevantes al dominio, utilizando las etiquetas *rdf:type* (clase), o *dcterms:subject* (categoría);
2. la extracción de información temporal, teniendo en cuenta la consistencia entre recursos "hermanos" (*owl:sameAs*) en distintas DBpedia;

ISSN 1135-5948

3. la integración de la información anterior en el corpus del dominio.

La identificación de fechas en un texto es una tarea relativamente compleja ante la ausencia de uniformidad y normalización para la incorporación de información temporal, en cualquier lengua, como el español (Vicente-Díez et al., 2008), (Vicente-Díez et al., 2010), (Vázquez Méndez y García Serrano, 2015) o el catalán (Llorens et al., 2009).

En este trabajo se identifican fechas tanto en español como en inglés. Se describe brevemente la estructura y la forma en que aparece la información temporal en DBpedia en la sección 1.1. El criterio sobre el que se identifican y extraen los recursos de la DBpedia relacionados con un dominio se basa en dos parámetros, el *tope* y la *profundidad*, y permite evitar la pérdida de precisión, al desechar *categorías* y *clases* no relevantes.

El corpus de dominio utilizado en este trabajo es el formado por las fichas semiestructuradas del corpus DIMH, que se describe con más detalle en la sección 1.2. La extracción de información temporal del corpus conlleva el tratamiento del multilingüismo (en oraciones cortas), el análisis de colisiones entre las fechas extraídas, y la representación de intervalos temporales (como se detalla en la sección 3). La notación inicial se realiza con TimeML (Pustejovsky et al., 2003), (Vázquez Méndez y García Serrano, 2015), referencia en la que se encuentra una descripción de otras herramientas. TimeML también se utiliza en el nuevo GATE-Time (Derczynski, 2016).

La información de cada ficha del corpus ha sido enriquecida con nuevas anotaciones de eventos relacionados con una línea de tiempo (alimentada automáticamente desde la DBpedia, según se describe en la sección 2), y de la información temporal extraída del corpus. Para ello se ha diseñado un formato de anotación temporal, *.moment* (descrito en la sección 3.4). El corpus enriquecido con la información temporal está disponible en un repositorio público (que se describe en la sección 4.1).

A pesar de que el modelo desarrollado no es dependiente del dominio, obviamente sí se aplica a dominios concretos, por lo que a la hora de realizar una interfaz que permita la experimentación de la propuesta y la evaluación futura de los expertos humanistas, se ha organizado la información léxica y temporal extraída siguiendo una aproximación no supervisada (García Serrano y Castellanos, 2016). Así, se hace posible la visualización del corpus anotado DIMH, tanto por su dimensión léxica como temporal (sección 4).

1.1 Organización de la DBpedia

En lo que sigue se describe brevemente la estructura de la información contenida en la DBpedia que es de interés en este trabajo.

La DBpedia (Lehmann et al., 2015) es un repositorio de datos etiquetados según varias ontologías. Su fuente original y principal de recursos es la Wikipedia, utilizándose principalmente métodos automáticos para la estructuración de los datos. Existen varios proyectos de DBpedia activos hoy día, cada uno conteniendo la información en un idioma específico, tal y como sucede con la Wikipedia. Por ejemplo en español es.dbpedia.org (Mihindukulasooriya et al., 2015).

Cada tema en la DBpedia es un "recurso". Como es de esperar, se tratan temas similares en cada DBpedia de idiomas distintos e incluso en otros recursos dentro de la misma DBpedia, creando lo que se denominan recursos "hermanos". Dado un recurso, podemos identificar a sus hermanos con la etiqueta *owl:sameAs*; así, el recurso (a), tiene como hermanos los recursos (b) y (c), siendo:
(a) es.dbpedia.org/page/Alzamiento_de_Varsovia
(b) it.dbpedia.org/resource/Rivolta_di_Varsavia/html
(c) de.dbpedia.org/page/Warschauer_Aufstand

Cada recurso contiene múltiples etiquetas que aportan la semántica a la información que engloban, por ejemplo: *rdf:type* (clase), o *dcterms:subject* (categoría). Además hay etiquetas que contienen fechas (con diferentes formatos), como: es.dbpedia.org/property/fecha o es.dbpedia.org/property/date.

Aunque no se puede garantizar que el valor contenido siempre cumpla con el formato ISO 8601[1], sí ocurre cuando se presenta el tipo *xsd:date*[2]. Sin embargo es difícil establecer cómo, porqué o para qué se ha definido una fecha en la DBpedia, ante la ausencia de unos criterios conocidos por todos al respecto.

[1] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=40874

[2] https://www.w3.org/TR/xmlschema11-2/#date

1.2 Corpus DIMH

El corpus DIMH (https://dimh.hypotheses.org/) consta de 7792 fichas que documentan con metadatos de material cartográfico imágenes de planos, mapas y dibujos del Archivo General de Simancas[3]. Cada ficha posee un identificador único y varias etiquetas con las que el anotador especificó los campos relativos a cada tipo de información incluida (por ejemplo <Titulo> o <Notas>, que incluyen texto libre). Estas fichas fueron enriquecidas con anotaciones relacionadas con entidades nombradas, sintagmas nominales y lemas (García-Serrano y Castellanos, 2016).

En este trabajo, se puede considerar a este corpus como de carácter general en tanto permite la experimentación en una tarea de extracción de información temporal.

2 Creación de una línea de tiempo

Para diseñar e implementar un método eficiente para alimentar una línea de tiempo de manera automática relacionada con un dominio en concreto, se decide utilizar la DBpedia (entre ellas la del español[4]). El método diseñado se divide en dos tareas: la de identificación y extracción de recursos acordes al tema deseado, y la de extracción de la información temporal contenida en estos recursos, para poder incluir eventos en la línea de tiempo, que tengan fecha de inicio y de fin.

2.1 Extracción automática de recursos de la DBpedia relevantes a un dominio

Se plantearon dos aproximaciones para abordar el problema, seleccionándose la segunda.

2.1.1 Modelo supervisado por un experto

En la DBpedia los recursos tienen asociados las etiquetas *clase*[5] (componente de la ontología que pertenece a la jerarquía, actualmente un grafo acíclico dirigido) y *categoría*[6] (agrupación de páginas que comparten un tema en común), siendo el conjunto de las *categorías* más amplio que el de las *clases*. Por ejemplo *Batalla de Lepanto* pertenece a la clase *Societal Event* y a más de una decena de categorías (*Batallas de España del siglo XVI, Batallas de la Armada de España, Guerras turco-venecianas, Reinado de Felipe II*…). Se seleccionó un conjunto de *clases* que acotan un conjunto de recursos relacionados con un tema del dominio, para utilizarlas como predicado en una consulta que ha de obtener los recursos necesarios para la identificación posterior de los eventos.

El problema es que el conjunto de clases que se obtienen automáticamente al inicio es muy pequeño; por tanto, en una primera fase se plantea expandir los recursos iniciales relacionándolos con nuevos recursos utilizando las *categorías* asociadas; luego se pondría a disposición de un experto humanista el conjunto de clases obtenido con esta expansión, y este elegiría el subconjunto con el cual se diseña el predicado para la búsqueda de recursos en la DBpedia.

A modo de ejemplo, partiendo del tema *Batalla de Lepanto*[7] vemos que posee la clase *MilitaryConflict*. De esta semilla se obtienen los recursos relacionados, con una consulta SPARQL como:

*SELECT * WHERE { ?s rdf:type dbpedia-owl:MilitaryConflict}*

Utilizando la propiedad *subject* de los recursos, se extraerán las categorías asociadas y con estas, más recursos con consultas SPARQL como:

*SELECT *WHERE { ?s dcterms:subject http://es.dbpedia.org/resource/Categoría:Batallas_de_España_del_siglo_XIII }*

Se entrará en un ciclo limitado por una cantidad de iteraciones (*profundidad*) cuya finalidad será obtener automáticamente un conjunto extenso de recursos del cual extraer las *clases*. Estas clases se presentan a un experto humanista para que escoja las adecuadas, por ejemplo *MilitaryConflict* y *MilitaryPerson*.

En una segunda fase se obtienen los recursos con la lista final de *clases* incluida en un predicado, con una consulta SPARQL como:

*SELECT * WHERE { {?s rdf:type dbpedia-owl:MilitaryConflict} UNION {?s rdf:type dbpedia-owl:MilitaryPerson}}*

Se han tenido en cuenta dos parámetros para las consultas a la DBpedia, *tope* (el límite a la lista de resultados que ha de devolver cada consulta de recursos de la DBpedia) y

[3] http://www.mcu.es/ccbae/es/mapas/principal.cmd
[4] http://es.dbpedia.org
[5] www.w3.org/1999/02/22-rdf-syntax-ns#type
[6] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms#subject
[7] http://es.dbpedia.org/page/Batalla_de_Lepanto

profundidad (la cantidad de iteraciones para expandir las categorías-recursos), sabiendo que incrementar su valor provoca que disminuya la calidad del resultado de la expansión de las *categorías*.

Para no perder precisión, se limitaron las *categorías* obtenidas en la segunda iteración y siguientes a aquellas que mostrasen una semejanza mínima (*similitud*) con las categorías obtenidas en la primera iteración. La medida utilizada expresa la relación entre el número de elementos del conjunto de recursos comunes contra el total definido por ambas:

$$S_{AB} = \frac{2 * \#recursos\ comunes}{\#recursos\ categoría\ A + \#recursos\ categoría\ B}$$

Por otra parte, como se observa que para nombrar las categorías en la DBpedia, se intenta utilizar un vocabulario reducido y expresiones sencillas, como por ejemplo Batallas_de_ Suecia_ del_siglo_XVIII, en vez de Acontecimientos_ bélicos_de_Suecia_del_siglo_XVIII, es posible filtrar automáticamente los términos que deben contener las *categorías* para ser incluidas. Para ello se utiliza una lista de términos aportada por los expertos humanistas denominada de conocimiento previo, porque representa la terminología de unos intereses o tarea concreta en un momento dado.

2.1.2 Modelo automático

Esta aproximación se basa en la anterior pero no utiliza la expansión categoría-recurso sólo como un método de obtención de nuevos recursos, sino también para obtener los recursos finales de los que extraer los eventos. Básicamente, se utilizan los recursos "comunes" entre las *categorías* que participan en el proceso de medición de similitudes a partir de la primera iteración, como salida final. Adicionalmente se aplicará el filtrado de términos del conocimiento previo utilizado en la aproximación anterior por los buenos resultados obtenidos.

2.1.3 Comparación de resultados

Debido a la no disponibilidad de un *gold standard*, la cobertura de los resultados plantea las dificultades que se detallan a continuación. Con la aproximación supervisada por expertos, los recursos para el resultado final se obtienen con una consulta, que utiliza el predicado escogido manualmente, obteniendo una cantidad de recursos imposible de evaluar manualmente. Entonces se establece un límite al número de resultados de la respuesta, pero aun así no está claro cómo establecer los parámetros *tope* y *profundidad* adecuados en la aproximación automática, para que la comparación entre ambas fuese efectiva.

Se decidió utilizar sólo la precisión como medida, porque evita la sensación de fracaso que provocaría a un usuario experto (humanista en el caso del corpus DIMH) el encontrar estos fallos en los resultados automáticos.

Se catalogaron como *falsos positivos* aquellos recursos obtenidos no acordes con el tema y *verdaderos positivos* a los acordes. Es interesante indicar que en el criterio para revisar manualmente estos últimos (del tema Guerras y Batallas), no sólo se incluyeron acontecimientos bélicos, sino personajes, cosas y lugares famosos exclusivamente por su participación en este tipo de eventos.

La precisión de la aproximación supervisada por expertos es de 0,9994 y la de la automática de 0,9476, luego ambas aproximaciones tienen una alta precisión. Se observó en las pruebas que la mayor causa de errores en la segunda aproximación está relacionada con personajes históricos no relacionados con el tema.

La primera aproximación es útil si la preselección de clases por un experto es necesaria o sencilla de llevar a cabo (que no lo es para un experto). Por lo tanto, en el experimento que se presenta, se ha decidido utilizar la aproximación automática ante una precisión tan cercana y así potenciar un proceso no supervisado por expertos.

2.2 Extracción de información temporal de recursos DBpedia

Para la detección de fechas que parecen dentro de los recursos DBpedia seleccionados, por ejemplo los identificados y descargados en la fase anterior, se han estudiado y comparado tres estrategias, seleccionándose la tercera.

2.2.1 Estrategia 1

Consiste en utilizar una herramienta de detección de información temporal en texto plano con HeidelTime (HT) (Strötgen y Gertz, 2010). Se desecha porque:

- Extraer fechas en un texto del que ya se ha realizado un etiquetado semántico, sería desaprovechar un trabajo previo para realizarlo de nuevo y con menor precisión.

- Detectar fechas no relevantes para el tema (actualizaciones del equipo de trabajo, etc.), genera un nuevo problema y por lo tanto para resolverlo habría que filtrar ese ruido.

2.2.2 Estrategia 2

Esta estrategia consiste en extraer primero información de ciertas etiquetas de la DBpedia de las que se conoce a priori que denotan fechas relacionadas con el tema. Las dos etiquetas que en la DBpedia en español que se utilizan al efecto son: http://es.dbpedia.org/property/fecha y http://es.dbpedia.org/property/date.

De un total de 23203 recursos que contienen estas etiquetas sólo 1036, un 4%, contienen un valor de tipo *xsd:date*. Una observación manual de un número grande (más de 150) de las propiedades "fecha", muestra que datan aspectos muy variados y sería necesaria la supervisión manual de cada propiedad para establecer las relevantes. Por ello se desecha esta estrategia, pues la hipótesis es potenciar una solución lo más desatendida posible.

2.2.3 Estrategia 3

Esta estrategia consiste en, primero obtener las propiedades de tipo *xsd:date*, que contienen las fechas en los recursos, y a continuación seleccionar las relevantes. Analizando los recursos que contienen esta propiedad se observa que las fechas no relevantes, denominadas "ruido", suelen referirse a meta-información sobre la creación del recurso, actualizaciones, etc. Sobre todo, se constata que las fechas relevantes suelen repetirse entre los recursos hermanos, mientras que las "ruido" no. Por lo tanto, como criterio, se utiliza la presencia de una fecha en varios recursos "hermanos".

Todos los recursos contienen una etiqueta *owl:sameAs* que define otros recursos (mayormente de DBPedia en otros idiomas) denominados recursos "hermanos", cuyo contenido es similar al que hay en español. De esta forma la estrategia seleccionada consiste en:

- Identificar todas las fechas en el recurso y en los recursos "hermanos", fechas definidas por la etiqueta *xsd:date*.
- Realizar un filtrado y dejar sólo las fechas que aparezcan en al menos 2 recursos.
- Escoger una fecha de inicio y de fin del intervalo del evento. De entre las fechas preseleccionadas se escoge la menor y la mayor, respectivamente.

3 Extracción de la información temporal en el corpus DIMH

Las fichas contienen información etiquetada aunque hay algunas zonas etiquetadas que no contienen información temporal relevante a la época histórica del objeto descrito, o de las entidades que se referencian (como las referencias bibliográficas) así que se desecharon.

3.1 Anotación inicial con HeidelTime

De acuerdo con la discusión sobre estándares y herramientas de anotación detallada en (Vázquez y García-Serrano, 2015), se ha escogido la herramienta HeidelTime[8] (HT), para la notación temporal inicial de las fichas del corpus DIMH. HT es un anotador temporal multilingüe para expresiones temporales, usando el estándar TIMEML (Pustejovsky, 2003) y crea marcas <TIMEX3> sobre el texto original que denotan expresiones temporales explícitas. Es configurable con opciones para especificar el dominio del texto y el idioma de procesado. Para este trabajo se creó un script que invoca a HT para cada una de las fichas y que crea un nuevo fichero con la información temporal anotada.

3.2 Representación en intervalos

La propuesta de representación está basada en la lógica temporal de Allen (1983), que define operadores para expresar las relaciones temporales entre los intervalos identificados en el texto (ficha). Por ejemplo X *During* Y, para indicar que el evento X sucede mientras el evento Y está sucediendo. La información temporal se representa en intervalos, porque expresa explícitamente su tiempo de creación/ interés/ pervivencia; la alternativa sería utilizar un punto, pero no sería posible expresar un mes de cierto año (1870-05) mientras que con un intervalo sí, el que tiene de inicio a (1870-05-01) y como fin a (1870-05-31).

Otro argumento a favor del intervalo es que facilita la resolución de relaciones temporales entre fechas y fichas. Si la ficha A tiene un evento con fecha 1850-05-25 y la ficha B con

[8] http://dbs.ifi.uni-heidelberg.de/index.php?id=129

(a) (b)

Figura 1: Interfaz de búsqueda sin resolución de colisiones (a) y con resolución (b)

1850-05 y se quisiese utilizar el sistema de puntos sustituyendo el día de mes faltante con el primer día del mes entonces A tendría 1850-05-25, B 1850-05-01 y no se podría establecer que los eventos se han solapado. Utilizando el sistema de intervalos, A (1850-05-25__1850-05-25) y B (1850-05-01__1850-05-31) sí se muestra el solapado, ambos eventos transcurrieron en mayo de 1850.

Los operadores utilizados para expresar las relaciones entre intervalos que permiten el análisis de colisiones son X *Before/ After/ Same/ MeetsBefore/ MeetsAfter/ Overlaps/ During* Y.

3.3 La extracción con resolución de colisiones

La información temporal en las fichas del corpus DIMH está presente en varios idiomas, predominando el español. HT permite procesar cada ficha en cada uno de los idiomas presentes en la ficha y dar por válida la información marcada si el idioma de procesamiento con HT coincide con el del texto marcado. Para implementar esta estrategia fue necesario el desarrollo de un módulo de detección de idiomas basado en modelos de Markov (Padró y Padró, 2014). Este módulo es independiente de dominio.

Una vez desarrollado el módulo anterior y realizada la anotación temporal del corpus DIMH, se observó que en una ficha se pueden encontrar varios intervalos que se contengan entre sí, además de períodos extremadamente extensos que se deben considerar en este dominio como "ruido". Como el objetivo de la extracción de esta información era anotar y agrupar las fichas en momentos históricos, y construir una interfaz de búsqueda en la dimensión léxica y temporal, se diseña e implementa una estrategia que, ante una colisión, selecciona el intervalo más específico. Por ejemplo, teniendo tres intervalos:

1733-01-01_1733-12-31 (año 1733)
1733-04-09_1733-04-09 (9 de abril de 1733)
1503-01-01_1805-12-31 (ruido)

y aplicando la resolución de colisiones sólo se extraería un intervalo, *1733-04-09_1733-04-09*.

El efecto en la interfaz gráfica es notable. Por ejemplo, cuando no se aplica la resolución de colisiones, hay fichas con grandes intervalos temporales sin relación con la obra referida, que crean una especie de "ventana temporal" muy amplia y que provoca una organización con relaciones entre conceptos con atributos temporales muy alejados en el tiempo. En el ejemplo de la figura 1 (a) se visualiza el concepto SigloXVIII-SigloXVII, que está relacionado con conceptos que contienen a su vez, los atributos SigloXVI, SigloXIX, SigloXX y SigloXXI.

Con la resolución de colisiones, los intervalos temporales suelen ser significativos (comprobación manual) y relacionados con la obra descrita en las fichas (plano, mapa o dibujo) (ver figura 1 (b)).

3.4 Integración

La integración de la información temporal del corpus DIMH y de los eventos extraídos de la DBpedia, se ha organizado alrededor de un conjunto de términos relevantes para los

expertos humanistas, denominado de conocimiento previo (Merás, 2016).

La información temporal del corpus DIMH obtenida, la línea de tiempo alimentada con eventos de la DBpedia y su información temporal asociada, se ha integrado utilizando un formato de anotación (ejemplo al final de esta sección), almacenado en el fichero *.moment* con cuatro secciones claramente definidas, una para la información temporal extraída <*time_temporalInformation*>, otra para los eventos relacionados de la línea de tiempo <*time_relatedEvents*>, otra para los términos del conocimiento previo encontrados <*existing_words*> y el resto para la información original. En un futuro, podrían estudiarse estándares del Linked Open Data (http://linkeddata.org/) y su posible aplicación.

```
<Ficha id="183679">
  <Tipo>Ilustraciones y Fotos</Tipo>
  <Titulo> [ Muestras de tripe común
labrado negro y felpa azul turquesa]
…
  <Publicacion>    <TIMEX3    tid="t1"
type="DATE" value="1776">1776</TIMEX3>
…
<time_temporalInformation>
<temporalInformationItem  start="1776-
01-01" end="1776-12-31" />
</time_temporalInformation>
<time_relatedEvents>
  <event id="0000000018" title= "Siglo
XVIII">
  <class> century</class>
  <moment  start="1701-01-01"  end  =
"1800-12-31" /> </event>
  <event    id="0000001059"    title=
"Reinado de Carlos III de España">
  <class> monarchs_europe </class>
  <class> monarchs_spain </class>
  <moment   start="1759-08-10"   end=
"1788-12-14" /> </event>
  </time_relatedEvents>
  <existing_words>
  <word id="0000000021" name="azul" />
  </existing_words>
</Ficha>
```

4 Interfaz para el análisis exploratorio

Para presentar a los expertos humanistas los resultados obtenidos y para realizar una búsqueda tanto por la dimensión temporal como por la dimensión léxica, se desarrolló una interfaz para la información organizada sobre la base de los términos del conocimiento previo, el contenido original de las fichas del corpus DIMH y la información temporal recabada, utilizando el modelo de visualización propuesto en (Filter, 2015).

Por ejemplo, si un experto comienza a buscar el término *fuerte*, cuando le aparecen los resultados podría filtrar adicionalmente si le interesa los del *siglo XVII* o los construidos con *Felipe II*; de manera contraria podría comenzar buscando a *Felipe II* como período histórico y luego decantarse por el término *fuertes*.

4.1 Recursos disponibles

Se puede acceder a la interfaz a través de la dirección: http://albali.lsi.uned.es:8111/ (ver figura 1 y figura 2). El repositorio https://github.com/meras0704/DBpediaTime contiene a:

(a) *Fichas_moment.zip*, que contiene 7792 ficheros en formato *.moment* con el resultado de la anotación y enriquecimiento de las fichas DIMH.

(b) *Solution_DBpediaTime*, solución implementada en Visual Studio y lenguaje C# con los programas necesarios. A su vez contiene 4 proyectos: *Blatella.Common*, la funcionalidad de uso común, *Blatella.ML.Common*, *TFM.Common*, la funcionalidad utilizada para la identificación de la información temporal, y *TFM.IU.WindowsForm*, la interfaz gráfica (primitiva pero suficiente).

5 Conclusiones y trabajos futuros

En este trabajo se ha propuesto un modelo experimental para la extracción de información temporal de la DBpedia y para enriquecer y contextualizar temporalmente la información de un corpus de textos de un dominio concreto. Por una parte se ha definido un modelo de extracción, basado en intervalos, que permite resolver las colisiones entre información temporal identificada en un corpus. Además se ha diseñado un método para la extracción automática de eventos de la DBpedia, integrándolos en una línea de tiempo, y se ha definido para ello el formato *.moment*.

Se ha experimentado la propuesta en el corpus DIMH y se ha desarrollado una interfaz de acceso a la información léxica y temporal.

Como trabajos futuros se desea confirmar la independencia del dominio del modelo propuesto para anotación temporal desde la DBpedia, plantear un modelo de evaluación no dependiente de los usuarios expertos y detectar patrones de evolución (Neouchi et al., 2001) en los atributos de una clase a lo largo del tiempo.

6 Agradecimientos

Este trabajo ha sido financiado parcialmente por los proyectos DIMH (HAR2012-31117) y Musacces (S2015/HUM3494).

Bibliografía

Allen, J. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11): 832-843.

Derczynski, L., J. Strötgen, D. Maynard, M. A. Greenwood and M. Jung. 2016. GATE-Time: Extraction of Temporal Expressions and Events. In *10th LREC*.

Filter J. 2015. Interactive Visualization of Large Concept Lattice. *Facultad de Ciencias de la Computación*, U. Magdeburgo. Alemania

Ganter B. 2002. Formal Concept Analysis: Methods and Applications. Computer Science. TU Dresden.

García-Serrano, A. y A. Castellanos. 2016. Conceptualización, acceso y visibilidad de la información en el proyecto DIMH. Cap. 16 *El dibujante ingeniero al servicio de la monarquía hispánica (XVI-XVIII)*, páginas 379-400. ISBN: 978-84-942695-6-1.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer and C. Bizer. 2015. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web J*ournal, 6(2): 167-195.

Llorens H., B. Navarro, E. Saquete. 2009. Detección de expresiones temporales TimeML en Catalán mediante roles semánticos y redes semánticas. *Procesamiento del Lenguaje Natural* (43): 13-21.

Merás A. 2016. Propuesta para extracción, representación y organización de información temporal en textos semiestructurados: aplicación al corpus DIMH. *Tesis del máster "Lenguajes y Sistemas Informáticos" de la UNED*.

Mihindukulasooriya N., M. Rico, R. García-Castro, A. Gómez-Pérez. 2015. An Analysis of the Quality Issues of the Properties Available. *Spanish Dbpedia, LNCS 9422*, páginas 198-209.

Neouchi R., A. Tawfik and R. Frost. 2001. Towards a Temporal Extension of Formal Concept Analysis. *Proceedings of the 14th Canadian Conference on AI*, Ottawa, Ontario.

Padró M., Ll. Padró. 2014. Comparing methods for language identification. *Procesamiento del Lenguaje Natural* (33): 155-161.

Pustejovsky J., J. Castaño, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer and G. Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in *Proceedings of the IWCS International Workshop on Computational Semantics*.

Strötgen, J. and M. Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, páginas 321-324, Uppsala, Sweden, July. ACL.

Tran, N., A. Ceroni, N. Kanhabua, and C. Niederée. 2015. Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. In *Proc. of the ACM International Conference on Web Search and Data Mining*, páginas 339-348.

Vázquez-Méndez, A. y A. García-Serrano. 2015. Anotación y representación temporal de tweets multilingües. *Procesamiento del Lenguaje Natural* (54): 53-60.

Vicente-Díez, M.T., D. Samy and P. Martínez. 2008. An empirical approach to a preliminary successful identification and resolution of temporal expressions in Spanish news corpora. *Proc. of the Sixth Int. Language Resources and Evaluation Conf. (LREC'08)*, Marrakech, Morocco, May, 2008, European Language Resources Association (ELRA), ISBN: 2-9517408-4-0, páginas 2153-2158.

Vicente-Díez M.T., J. Moreno-Schneider, P. Martínez. 2010. Temporal information needs in ResPubliQA: an attempt to improve accuracy. The UC3M Participation at CLEF 2010, *CLEF 2010 LABs and Workshops, Notebook Papers*, Padova, Italy, September.

Zhang, L., W. Chen, T. Tran and A. Rettinger. 2015. Time-Aware Entity Search in DBpedia. In *European Semantic Web Conference, p*áginas175-179.