Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 317-318 recibido 30-04-2011 aceptado 24-05-2011

MDFaces: An intelligent system to recognize significant terms in texts from different domains using Freebase[1]

MDFaces: Un sistema inteligente para reconocer conceptos relevantes en textos de diferentes dominios usando Freebase

Fernando Aparicio, Rafael Muñoz, Manuel de Buenaga, Enrique Puertas
Departamento de Sistemas Informáticos y Automática, Escuela Politécnica
Universidad Europea de Madrid
C/ Tajo, s/n. Villaviciosa de Odón.
{fernando.aparicio,rafael.munoz,buenaga,enrique.puertas@uem.es}

Resumen: *MDFaces* (Multi-Domain Faces) es un sistema inteligente capaz de reconocer conceptos relevantes en textos, procedentes de diferentes dominios, y mostrar información detallada y semántica relacionada con estos conceptos. Para su desarrollo se hace uso de una metodología que utiliza datos procedentes de la ontología de conocimiento general llamada *Freebase*. En particular, se ha implementado esta metodología para los dominios médico y turístico.
Palabras clave: Reconocimiento de entidades nombradas, sistemas basados en conocimiento, ontologías, recuperación de información.

Abstract: *MDFaces* (Multi-Domain Faces) is an intelligent system that allows recognition of relevant concepts in texts, from different domains, and shows detailed and semantics information related to these concepts. For its development, it is have been employed a methodology that uses a general knowledge ontology called *Freebase*. In particular, we have implemented this methodology for medical and tourism domains.
Keywords: NER, knowledge based systems, ontologies, information retrieval.

1 Introducción

Las ontologías son un recurso cada vez más utilizado para almacenar conocimiento en diferentes disciplinas, existiendo una gran variedad y heterogeneidad de fuentes y formatos. A pesar de los enormes esfuerzos realizados para la homogeneización y estandarización de las mismas, siguen siendo un recurso con un alto grado de dependencia del dominio en la mayoría de los casos. Una de sus aplicaciones, en las tareas de procesamiento del lenguaje natural, es la anotación de textos, como se puede encontrar en (Jonquet, Shah y Musen, 2009) para el ámbito médico.

En este trabajo presentamos un sistema Web basado en la aplicación a dos dominios distintos de una metodología que, partiendo del conocimiento almacenado en Freebase (Bollacker et al., 2008), permite, por un lado, extraer listas de conceptos para realizar tareas de reconocimiento de entidades nombradas en un texto origen y, por otro lado, obtener relaciones semánticas e información descriptiva vinculada con los conceptos detectados.

El esquema básico de los datos de Freebase está basado en los denominados *dominios*, *tipos*, *tópicos* y *propiedades*. Los tópicos mantienen una relación del tipo "es un" con los tipos y "tiene un" con las propiedades, que contienen información agrupada bajo otros tipos. Uno de los mecanismos de acceso a esta ontología es el *Metaweb Query Language* (Meyer et al., 2010), que admite el acceso a través del protocolo HTTP y puede ser utilizado en procesos M2M y, por tanto, ser ejecutado bajo un servidor Web.

[1] Este trabajo ha sido financiado por los proyectos MEDICAL-MINER (TIN-2009-14057-C03-01) y MA2VICMR (S2009/TIC-1542).

2 Descripción

La metodología propuesta se divide en dos etapas bien diferenciadas, que se describen a continuación.

En primer lugar, es necesario seleccionar un dominio en Freebase y estudiar los tipos de interés que agrupa, escogiéndose aquellos que dispongan de un número suficientemente significativo de tópicos. A continuación se construye una consulta MQL para cada uno de los tipos elegidos (es decir, una consulta genérica como la que se muestra en la Figura 1), con la que se descargan las listas de conceptos (nombres de los tópicos) utilizadas para el reconocimiento de entidades nombradas. Este reconocimiento se encuentra integrado en la lógica del servidor y se lleva a cabo con el componente Gazetteer del sistema ANNIE incluido en el software GATE[2]. En nuestro caso, se han seleccionado los dominios médico (*Medicine* en Freebase) y turístico (*Travel* en Freebase).

```
[{
 "type": TIPO_DE_FREEBASE, "name~=":"*",
 "id":null, "name":null, "optional":false, "limit":20000
}]
```

Figura 1: MQL para la recuperación de las listas de nombres de tópicos

En la segunda etapa se estudia la información detallada (datos sobre el tópico como una descripción e imágenes) y la información semántica (propiedades asociadas al tópico) que se pretende incluir en el sistema. Si entre los tipos seleccionados existe una relación circular (si las propiedades de los tópicos agrupados bajo uno de los tipos pueden ser obtenidas a partir de tópicos agrupados bajo los otros tipos), entonces toda la información semántica relacionada con los conceptos se puede mostrar a través del sistema desarrollado, mientras que si no existe esta relación se puede enlazar directamente al contenido en la Web de Freebase. Para obtener esta información se hace uso de consultas MQL con las que se solicita propiedades y datos de un tópico. En nuestro caso se han seleccionado los dominios "medicine" y "travel". En el dominio médico los tipos seleccionados son *Disease or medical condition*, *Medical treatment* y *Symptom*. En el dominio turístico los tipos seleccionados son *Tourist attraction*, *Accommodation* y *Travel destination*. En ambos casos para cada concepto se obtiene la siguiente información relacionada: el artículo, las imágenes y un conjunto de propiedades, como por ejemplo los síntomas, tratamientos y factores de riesgo asociados a una enfermedad (en el caso del tipo *Disease or medical condition* del dominio médico) o el alojamiento y las atracciones turísticas relacionadas con un destino (en el caso del tipo *Travel destination* del dominio turístico).

Como ejemplo, en la Figura 2 se muestra el resultado tras el reconocimiento de entidades médicas en un fragmento de historial clínico. En función del origen de la detección, la información obtenida al seleccionar estos conceptos procederá de los servicios de Freebase o de Medlineplus.

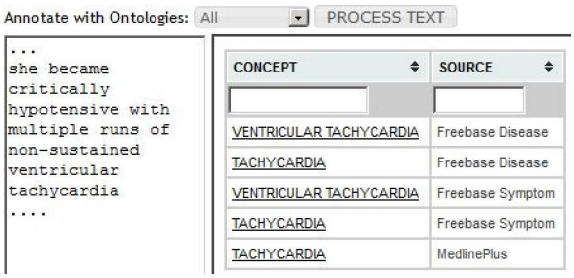

Figura 2: Resultado inicial del interfaz Web al procesar un fragmento de historial clínico.

Referencias bibliográficas

Bollacker, K., C. Evans, P. Paritosh, T. Sturge y J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. En *Proceedings of the international conference on Management of data, ACM SIGMOD, pp.* 1247-1250. *Vancouver (Canada*).

Jonquet, C., N. Shah y M. Musen. 2009. The Open Biomedical Annotator. En *AMIA Summit on Translational Bioinformatics,* pp. 56-60, *San Francisco (CA)*.

Meyer S., J. Degener, J. Giannandrea y B. Michener. 2010. Optimizing schema-last tuple-store queries in graphd. En *Proceedings of the International Conference on Management of Data, ACM SIGMOD,* pp. 1047-1056, *Indianapolis (Indiana)*.

[2] http://gate.ac.uk