Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 213-216 recibido 08-05-12 revisado 31-05-12 aceptado 05-06-12

Servicios de anotación y búsqueda para corpus multimedia

Annotation and Search Services for Multimedia Corpus

David Hernández-Aranda
NLP&IR Research Group
ETSI Informática, UNED,
Madrid, Spain
daherar@lsi.uned.es

Rubén Granados
NLP&IR Research Group
ETSI Informática, UNED,
Madrid, Spain
rgranados@lsi.uned.es

Ana García Serrano
NLP&IR Research Group
ETSI Informática, UNED,
Madrid, Spain
agarcia@lsi.uned.es

Resumen: En este artículo corto se muestra la funcionalidad tanto del servicio anotador de textos desarrollado en el marco del proyecto Buscamedia[1], como del buscador sobre recursos o documentos multimedia anotados.
Palabras clave: Recuperación de información multimedia, Anotación multimedia, Recuperación de información textual, Fusión multimedia, Corpus.

Abstract: This paper shows the textual annotator service developed in the project Buscamedia as well as the search performed on multimedia resources or documents annotated.
Keywords: Multimedia information retrieval, Multimedia Annotation, Text-based Information Retrieval, Multimedia Fusion, Corpus.

1 Introducción

La recuperación de información multimedia (texto, imágenes, audio, vídeo) se aborda con enfoques textuales en la mayoría de las herramientas y sistemas existentes, usando anotaciones y metadatos asociados a las imágenes (Depeursinge and Müller, 2010), al audio o a los vídeos (Hernández-Aranda et al., 2010), o una parte de ellos, como son los segmentos, las instantáneas o *keyframes*, etc (Geurts et al 2005). Por ello, la anotación automática de recursos multimedia, sin intervención humana, está en continua investigación (Feng and Lapata, 2010), (Lombardo and Damiano 2012).

Sin embargo en el proyecto español Buscamedia se afronta el problema con una aproximación netamente multimedia, para lo que se han desarrollado subsistemas que "entienden" y procesan los recursos multimedia como se presenten (identificando por ejemplo los personajes que intervienen, objetos físicos, etc.). Cuando el resultado del análisis de estos subsistemas son anotaciones en forma de texto, se integran en el subsistema textual.

En las secciones siguientes, se presenta brevemente el sistema desarrollado y cómo buscar en el corpus *Deportes20*. A continuación se describe el corpus desarrollado y sus anotaciones provenientes del análisis de los recursos multimedia, como son las transcripciones, los subtítulos, algunos objetos físicos en imágenes, el texto sobreimpreso, los logos y las moscas. Se sigue con una breve presentación del servicio anotador textual y finalmente se muestran algunos ejemplos de búsqueda orientados a la validación del sistema a través de una prueba de concepto.

2 Descripción del sistema de búsqueda

El prototipo desarrollado consta de una interfaz web que permite la búsqueda y la visualización de resultados a partir de una consulta dada, siguiendo las pautas de un buscador común, pero que además permite mostrar todas las funcionalidades desarrolladas ya que los "botones" del interfaz las representan.

La visualización de los resultados se realiza a partir de los *snippets* creados manualmente sobre archivos multimedia de los vídeos o textos (noticias o páginas web) del corpus

[1] http://www.cenitbuscamedia.es/

Deportes20, como pueden ser imágenes o *keyframes* y segmentos de visión con información concreta, o textos multilingües porque los documentos multimedia están en castellano, catalán, euskera o inglés.

Figura 1: Interfaz del sistema de búsqueda

En la figura 1 se muestra el interfaz correspondiente al sistema de búsqueda textual en el que se han realizado los siguientes pasos:

Preproceso. Con la información textual extraída de los documentos multimedia, se realiza un análisis de detección de las entidades nombradas. Dependiendo del idioma, se aplica una herramienta distinta. Para el castellano e inglés, *Stilus* (licencia para investigación proporcionada por Daedalus[2]), y para el catalán, se aplica *Freeling*[3].

Para el caso de los vídeos en catalán, se hace uso además de un recurso externo, el denominado *Thesaure* (con licencia de uso restringida de la corporación catalana de televisión), con el objetivo de enriquecer la anotación en la que aparezcan términos de dicho tesauro.

A continuación se crea un documento único para cada documento multimedia unificando la información contenida originalmente, y además se añaden el idioma, el nombre del documento original, y las entidades nombradas detectadas, o los campos relacionados con la información semántica del tesauro.

Un ejemplo de documento XML único es:

```
<out>
   <idioma>es</idioma>
   <subtitulos>Y el líder en Liga sigue
      siendo el Real Madrid.. Los
      blancos debutan esta semana en
      Copa...</subtitulos>
   <textoSobreimpreso>KOREAN PETRONAS
      Formula1. ronaldo 7. P.LEÓN. The
      next big Audi. ¿Te llevo?. LA
      NOCHE DE CR7. Audi. RONALDO...
      </textoSobreimpreso>
   <logos>PETRONAS Bwin Audi Mahou
      Adidas mahou audi Audi AUDI
      bwinCamiseta Real Madrid C.F. LFP
      bwin RNE Punto pelota Onda cero
      Barça TV TV...</logos>
   <nes> Real_Madrid_Club_de_Fútbol
      MADRID Madrid REAL Real_Madrid
      Comunidad_Autónoma_de_Murcia
      Murcia José_Mourinho Mourinho
      Mou...</nes>
   </out>
```

A partir de estos documentos únicos el pre-procesamiento sigue con los analizadores *SnowBall* implementados para cada idioma en *Lucene*, para efectuar el *stemming*, y con la eliminación de stopwords.

Indexación. El modelo de indexación consiste en la creación de un único índice haciendo uso de *Lucene*, indexando en diferentes campos toda la información de los cuatro idiomas.

Búsqueda. Una vez indexado el corpus *Deportes20*, en la búsqueda de cada consulta se obtendrá una única lista de resultados ordenados por relevancia. La función de *ranking* utilizada es BM25F que extiende a BM25 para documentos estructurados (formados por campos).

En este prototipo se permite la selección del operador con el que se desea hacer la búsqueda (OR o AND). Además se podrán seleccionar los tipos de metadatos, correspondientes a los distintos tipos de anotaciones, del documento único. Y se podrán filtrar los resultados recuperados por el tipo de documento (solo vídeos, solo noticias/páginas web o ambos).

El servicio de búsqueda está disponible para otros investigadores, y previa solicitud de *login* y *password* pueden acceder al prototipo en la dirección siguiente: http://albali.lsi.uned.es/deportes20-1.0.0/.

3 *El corpus Deportes20*

La colección está compuesta por 4 tipos de recursos o documentos multimedia:

[2] www.daedalus.es
[3] http://nlp.lsi.upc.edu/freeling/

Vídeos en catalán (proporcionados por CCMA[4], miembro del consorcio): 21 documentos multimedia en catalán, de los cuales, 10 tienen asociado un documento XML con su descripción, una carpeta con *keyframes* asociados, y los objetos detectados que aparecen en ellos.

De los 11 recursos restantes sí que se dispone de sus vídeos correspondientes, así como de sus transcripciones y *keyframes* asociados. Sin embargo, en este caso, no se dispone de los objetos que aparecen en ellos.

Vídeos en castellano (proporcionados por ISID[5], miembro del consorcio). 10 vídeos en castellano, de los cuales 6 tienen asociado un vídeo, la transcripción del audio y los subtítulos de dicho vídeo. Los 4 vídeos restantes, además de la información anterior, contienen el texto sobreimpreso y los logos y moscas detectados en cada vídeo.

Páginas Web (proporcionadas por Daedalus, miembro del consorcio). Son 34 páginas web en formato HTML, cuya temática está relacionada con los vídeos de los grupos anteriores, de las cuales 30 están en idioma castellano, 3 en catalán y 1 en inglés (selección manual).

Noticias (proporcionadas por *Daedalus*). Conjunto de 62 noticias en formato HTML, de las cuales 30 están en castellano, 30 en catalán y 2 en euskera. Se extrajeron con consultas relacionadas con los documentos de un corpus de 21.632 noticias de 16 periódicos con formatos diferentes.

Con todo ello, se construye una colección anotada de 127 recursos o documentos multimedia correspondientes a vídeos, páginas web y noticias textuales. Este corpus está disponible para la comunidad de investigadores, previa solicitud.

4 Anotacion automática

La herramienta de anotación textual desarrollada en el proyecto permite analizar textos en diferentes idiomas (español, inglés, catalán) y realizar su análisis morfosintáctico, de forma que se obtengan los términos que pertenecen a una categoría morfosintáctica específica, y las entidades nombradas.

Para ello, esta herramienta utiliza módulos intermedios que sirven de *wrapper* para herramientas conocidas, como son *FreeLing*, *TreeTagger*[6], *Stanford NER*[7] y *Stilus* de *Daedalus*, y que serán seleccionados en la llamada al servicio. El servicio web de esta herramienta de anotación se encuentra en: http://albali.lsi.uned.es/DemoAnotadorWS/ y puede utilizarse para investigación, previa petición de *login* y *password*.

En la figura 2 puede observarse una salida del interfaz de este servicio.

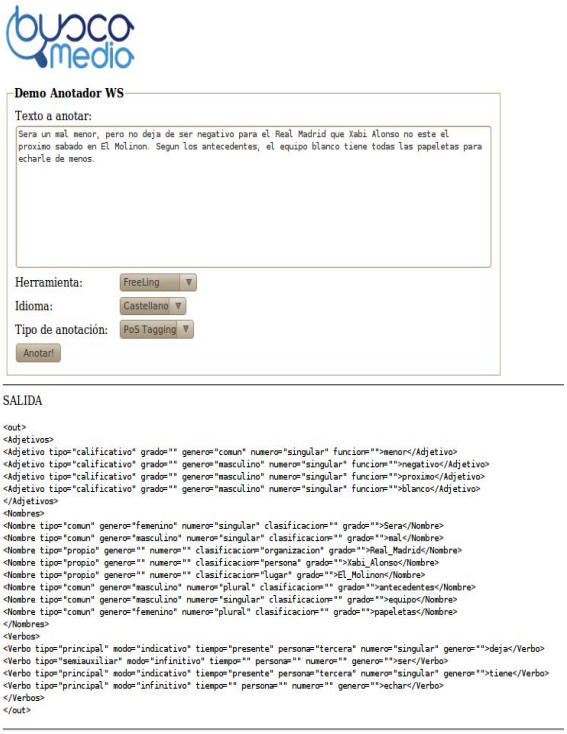

Figura 2: Interfaz del anotador

5 *Pruebas con la colección Deportes20*

El corpus *Deportes20* se complementa con un conjunto de consultas (relacionadas con el objetivo a evaluar) y sus juicios de relevancia (la mayoría con explicaciones detalladas).

A continuación se incluyen algunos ejemplos de prueba, indicando en cada uno de ellos las ventajas alcanzadas con la combinación de anotaciones provenientes de diferentes medias.

Consulta 1: "uso de un zeppelín en eventos deportivos".

Opciones de búsqueda: castellano, OR, Vídeos, objetos

Resultados: En este ejemplo el usuario busca vídeos sobre la aparición de un zeppelín

[4] http://www.ccma.cat/pccrtv/ccrtvSeccio.jsp

[5] www.isid.es

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[7] http://nlp.stanford.edu/software/CRF-NER.shtml

en algún evento deportivo, donde estos suelen usarse con fines publicitarios. Para recuperar el único vídeo del corpus relevante para esta consulta será necesario hacer uso del campo "Objetos", ya que el vídeo buscado está anotado por el servicio de anotación de objetos con la aparición de un zeppelín.

Con este ejemplo queda clara la utilidad de la integración de la salida proveniente del servicio de detección de objetos en vídeos, en forma de información textual. Anotando textualmente el objeto multimedia con la información sobre la identificación del objeto "zeppelín", se consigue posteriormente recuperar el vídeo buscado (correspondiente a un partido de baloncesto en el que aparece un zeppelín).

Figura 3: Interfaz del buscador para la respuesta a la consulta "Cristiano Ronaldo"

Consulta 2: "Información sobre el Casademont".

Opciones de búsqueda: castellano, OR, Vídeos, Thesaure>Detalles

Resultados: La consulta busca vídeos relacionados con un equipo de baloncesto, el Casademont. Si no se hace uso de la información semántica que proporciona el Thesauro (en Detalles), nunca se sabría que se refiere al equipo que actualmente se llama Akasvayu Girona (antiguamente Casademont Girona). Por lo tanto, gracias a la selección de la opción "Detalles" del Thesaure, la búsqueda recupera el vídeo del corpus que trata sobre un partido entre el Barcelona y el Girona.

6 Comentarios finales

Se ha presentado el servicio de anotación de recursos multimedia y unos ejemplos de búsqueda que muestran los beneficios de la anotación.

A continuación se pretende abordar la integración semántica de las anotaciones a través de la información contenida en una ontología multimedia disponible en el proyecto Buscamedia.

7 Agradecimientos

Este trabajo se ha financiado con el proyecto competitivo BUSCAMEDIA (CEN-20091026), financiado por el Ministerio de Industria.

Agradecemos muy especialmente la colaboración de los investigadores de todos los miembros del consorcio, pero muy en particular en esta tarea de creación del corpus a los de *Tecnalia*, *UC3M*, *ISID*, *Bilbomática*, y por supuesto a *Daedalus* y *ATOS*.

Bibliografía

A. Depeursinge and H. Müller, (2010). *Fusion Techniques for Combining Textual and Visual Information Retrieval*. In H. Müller, P. Clough, T. Deselaers, & B. Caputo, Experimental Evaluation in Visual Information Retrieval (Vol. 32). Springer.

A. García-Serrano, R. Granados, D. Hernández-Aranda, V. Fresno y J. Cigarrán, *Anotación para la recuperación de información multimedia: el corpus Deportes20*, Actas del congreso SEPLN 2012, Valencia, 2012.

J. Geurts, J. van Ossenbruggen, L. Hardman, *Requirements for practical multimedia annotation*, Workshop on Multimedia and the Semantic Web, 4-11 2005.

D. Hernández-Aranda, R. Granados, J. Cigarran, A. Rodrigo, V. Fresno, and A. García-Serrano. *UNED at mediaeval 2010: exploiting text metadata for automatic video tagging*. In MediaEval 2010 Workshop. Pisa, Italy, 24 October, 2010.

V. Lombardo and R. Damiano, *Semantic annotation of narrative media objects*, Multimedia Tools and Applications, Volume 59, Number 2, Pages 407-439, 2012.

Y. Feng and M. Lapata. 2010. *Topic Models for Image Annotation and Text Illustration*. In Proc. of the Human Language Technologies at the North American Chapter of the Association for Computational Linguistics, 831-839. Los Angeles, California.