ANTRO: un sistema de reconocimiento y gestión de antropónimos

Daniel Casanova, Xavier Lloré, Rafael Marín, Josep M. Merenciano, Gema Pérez, David Trotzig
{dcasanova, xllore, rmarin, jmmerenciano, gperez, dtrotzig}@planeta-actimedia.es

Departamento de Lingüística Computacional
Banco de Contenidos, Planeta Actimedia, S.A.

Resumen. En este artículo se describe ANTRO, una herramienta que permite identificar los antropónimos susceptibles de aparecer en cualquier tipo de texto, así como gestionarlos adecuadamente dentro de un sistema general de procesamiento del lenguaje natural.

1. Introducción

El tratamiento de los nombres propios supone un reto interesante para cualquier sistema de procesamiento del lenguaje natural. Ello se debe, en buena medida, al hecho de que constituyen un conjunto de elementos difícilmente acotable y con un grado de variabilidad elevado.

Ahora bien, las ventajas que se derivan de un adecuado análisis de los nombres propios son claras, ya que pueden influir en la mejora de procedimientos tales como la extracción y recuperación de información o la clasificación temática de documentos.

En ámbitos más específicos como, pongamos por caso, el mundo editorial, las posibles utilidades también se hacen evidentes: creación de índices onomásticos, sistematización en el uso de nombres propios, etc.

2. Arquitectura del sistema

ANTRO se inserta dentro de un sistema de procesamiento del lenguaje natural cuya pieza clave es una base de conocimiento (BDCon) sobre la que pivotan una serie de herramientas lingüísticas. La puerta de acceso a la BDCon es un diccionario de lemas (DICO), lo cual exige someter los textos a un proceso previo de lematización. La entrada de ANTRO es tambien la salida de este proceso: una secuencia de lemas etiquetados morfológicamente.

3. Obtención de los datos

Como es sabido, los nombres propios —y, en especial, los antropónimos— no aparecen siempre en los textos con la misma forma superficial. Así, para referirnos a *Federico García Lorca*, podemos utilizar, además de la anterior, varias alternativas: *F. García Lorca*, *García Lorca* e incluso *Lorca*.

Con este objetivo en mente, hemos diseñado un sistema que permita almacenar y gestionar las diferentes variantes con que puede aparecer un antropónimo en un texto.

El procedimiento de generación de variantes se ha implementado en dos módulos. El primero es una gramática de cláusulas definidas que reconoce los distintos formatos de cita e identifica sus constituyentes. El segundo puede verse como un traductor que aplica a sus elementos todas las posibles operaciones de generación. Las transformaciones generales para la gran mayoría de antropónimos generan tres variantes principales: a) sólo el nombre, b) sólo el/los apellido/s y c) nombre y apellidos. En todos los casos, hemos establecido una Forma de Referencia (FR) que, por lo general, coincide con la variante que contiene mayor información.

De este modo, a partir de una FR como *Roald Engelbert Amundsen* se crean las siguientes variantes:

Forma	**TipoVar**
Roald Engelbert Amundsen	FR
R. Engelbert Amundsen	var
Roald E. Amundsen	var
R. E. Amundsen	var
R.E. Amundsen	var
Roald Engelbert	var
Amundsen	var

Hemos adoptado como punto de partida una base de datos con 61.432 antropónimos (correspondientes a las entradas biográficas de una enciclopedia) que, al aplicarles este proceso, se han convertido en 214.215 variantes

4. Gestión de los datos

Toda esta información debe estructurarse de forma tal que permita una gestión y un mantenimiento adecuados y sea consistente con el sistema general. Los puntos que deben tomarse en consideración son: (1) una misma entidad de la BDCon puede tener distintas formas; (2) hay información específica de cada forma e información específica del conjunto de variantes con el mismo referente; (3) el referente de un conjunto de variantes con la misma forma puede no ser único.

Para facilitar las búsquedas y por coherencia con las entradas léxicas de DICO, cada forma o variante se considera una entrada (un lema) del diccionario. Dichos lemas se agrupan en conjuntos de lemas que llamamos **metaformas**. Inicialmente todas las variantes que se han generado a partir de una misma forma (ver apartado anterior) se han asociado a una misma metaforma. No obstante, el sistema permite que la agrupación de lemas en metaformas pueda realizarse por cualquier criterio; es más, se permite la posibilidad de que una misma forma pueda pertenecer a distintas metaformas al usar criterios de agrupación distintos. Obsérvese que el criterio inicial de generación de metaformas supone que una misma variante obtenida de formas distintas aparece una sola vez como lema, pero pertenece a distintas metaformas.

El concepto de metaforma resuelve el requisito (1), y lo hace sin la necesidad de generar lemas repetidos para los casos de homonimia.

El hecho de poder asociar atributos tanto en el lema como en la metaforma resuelve el requisito (2). Dentro de todas las variantes pertenecientes a una metaforma, una de ellas se marca como forma de referencia (FR); ello permite usarla como elemento representativo del conjunto (y puede ser la base de la regularización de textos).

El requisito (3) se resuelve con una relación 1:n entre las metaformas y la BDCon. Así, el acceso a la BDCon no se realiza directamente a través de DICO, sino a partir de las metaformas.

La necesidad del requisito (3) se ve más clara con un ejemplo. Como es sabido, "Barcelona" puede hacer referencia a la Ciudad Condal o a una ciudad de Venezuela. Si como variantes se considera sólo "Barcelona" se podría decidir tener una sola metaforma con ambigüedad referencial (aunque tambien se podrían definir dos metaformas). Por el contrario, si aceptamos las variantes "Barna" y "Ciudad Condal", que sólo son aplicables a una de las dos entidades, necesariamente tendremos dos metaformas distintas.

5. Descripción de la herramienta

En su estado actual de desarrollo, ANTRO ofrece dos tipos de funcionalidades básicas: las de consulta y las de análisis. Las consultas permiten, además de saber si un antropónimo (ya sea como FR o como variante) figura en la base de datos, obtener información sobre las variantes relacionadas con un antropónimo determinado. Téngase en cuenta que las consultas pueden realizarse directamente a través del teclado o bien a través del análisis de un texto.

Por otra parte, la función principal del análisis consiste en reconocer los antropónimos que aparecen en un texto. Posteriormente, cada una de las formas identificadas se marcan en función de los atributos correspondientes a dicha forma (e.g. si es una FR, una variante o un NP desconocido).