Programa deductor de elementos morfológicos en contextos de oraciones infinitas iterativas

Carlos Alonso Hidalgo Alfageme
Universidad Autónoma de Madrid
C/ Mota del Cuervo 8, 8º 4. Madrid 28043
alonsohidalgo@yahoo.com

Resumen: El programa deduce elementos morfológicos en contextos formados por oraciones infinitas iterativas no necesariamente segmentadas en palabras. Experimentos con contextos pequeños prueban que en ellos la forma de las unidades morfológicas es variable. El programa no utiliza información morfológica contenida en bases de datos, lo que permite el análisis de contextos multilingües.
Palabras clave: morfología, segmentación, contexto, multilingüe

Abstract: The program deduces morphologic elements in contexts formed by not necessarily segmented in words iterative infinite sentences. Experiments with small contexts prove that in them morphologic units' form varies. The program does not use morphologic information extracted from databases, which allows analyzing multilingual contexts.
Keywords: morphology, segmentation, context, multilingual

1 Presentación del programa

A veces es necesario realizar investigación morfológica base y dedicar esfuerzos al estudio de oraciones posibles pero no probables. La aplicación informática que presentamos analiza la morfología de los elementos de oraciones infinitas regulares sujetos a un esquema sintáctico determinado.

En la primera fase del análisis se produce una segmentación de la oración propiciada, precisamente, por la regularidad con que en ella se repiten cadenas de signos. En la segunda fase del análisis, la comparación de las cadenas obtenidas de diferentes oraciones destila lexemas y morfemas. El algoritmo utiliza sólo información obtenida del objeto de análisis, por lo que demuestra fragmentar correctamente oraciones codificadas en diferentes idiomas.

2 Primera fase del análisis

La aplicación informática opera con cadenas infinitas de signos que no necesariamente han de presentarse segmentadas en unidades menores. Las cadenas corresponden al esquema de la siguiente,

(1) unacajacontieneunacajaquecontiene unacajaquecontieneunacajaquecon...

donde una primera cadena irregular (unacaja) precede a la cadena que se repite hasta el infinito (contieneunacajaque):

(1) unacaja / contieneunacajaque / contieneunacajaque / ...

La primera cadena irregular se usa como plantilla que permite la división en tres segmentos de la cadena que se repite:

(1) unacaja / contiene unacaja que / contiene unacaja que / ...

3 Segunda fase del análisis

En la segunda fase del análisis la oración estudiada se coloca en un contexto matricial de oraciones paralelas comparables en el que los segmentos diferentes son intercambiables en una misma columna (en negrita) sin que las oraciones situadas en las filas pierdan su gramaticalidad:

(2)	**lacaja**	/ contiene	/ **lacaja**	/ que	/ contiene	/ ...
(3)	**micaja**	/ contiene	/ **micaja**	/ que	/ contiene	/ ...
(1)	**unacaja**	/ contiene	/ **unacaja**	/ que	/ contiene	/ ...
(4)	**unabonitacaja**	/ contiene	/ **unabonitacaja**	/ que	/ contiene	/ ...

Contexto α

En las columnas de la matriz se aplican tres reglas:

(a) Si en una columna todos los segmentos presentan signos comunes ordenados desde el límite izquierdo de los segmentos, entonces es posible marcar una división entre el conjunto de signos comunes ordenados y el resto de los segmentos.

(b) Si en una columna todos los segmentos presentan signos comunes ordenados desde el límite derecho de los segmentos, entonces es posible marcar una división entre el conjunto de signos comunes ordenados y el resto de los segmentos.

(c) Si el conjunto de todos los signos ordenados de un segmento está contenido en otro segmento de la misma columna, alcanzando el conjunto el límite derecho del segmento mayor, entonces es posible marcar una división entre el conjunto de signos comunes ordenados y el resto del segmento mayor.

Estas tres reglas establecen en el marco de una columna tanto la frontera entre palabras como la frontera entre los morfemas de la palabra. La regla (b) es la que divide en el contexto α el sustantivo de los adjetivos. La misma regla distingue el prefijo de la raíz verbal en el contexto β:

(5) unacaja / **de-tiene** / unacaja / que / **de-tiene** / ...
(6) unacaja / **re-tiene** / unacaja / que / **re-tiene** / ...
(1) unacaja / **con-tiene** / unacaja / que / **con-tiene** / ...

Contexto β

Las tres reglas se pueden aplicar en cualquier orden y de modo recursivo. En un principio se consideró un número muy superior de reglas que se fueron descartando paulatinamente hasta la obtención de las tres que en todo caso (1) no llevan el proceso de división más allá de la obtención de morfemas y (2) no detienen el proceso de división antes de la obtención de morfemas. No superó esta fase experimental previa, por ejemplo, la variante izquierda de la regla (c).

La importancia teórica de estas tres reglas no reside en el amplio porcentaje de elementos que demuestran aislar correctamente, sino en aquellos casos en los que la división que las reglas ofrecen no es la que cabría esperar. El contexto β' añade una oración al contexto β:

(5) unacaja / **detiene** / unacaja / que / **detiene** / ...
(6) unacaja / **retiene** / unacaja / que / **retiene** / ...
(1) unacaja / **contiene** / unacaja / que / **contiene** / ...
(7) unacaja / **enfría** / unacaja / que / **enfría** / ...

Contexto β'

En el contexto β', al contrario de lo que sucedía en el contexto β, no hay ninguna razón para realizar las divisiones "**de-tiene**", "**re-tiene**" y "**con-tiene**", cuya pertinencia, por otra parte, continúa pareciendo evidente. La realización de tales divisiones, con la consiguiente violación de las reglas que hemos presentado, justificaría también la siguiente división errónea de los elementos de una hipotética columna: "**con-juga**", "**con-struye**", "**con-ciencia**", "**enfría**".

Tras una larga serie de experimentos con reglas coherentes llegamos a la conclusión de que lo que se considera *morfema* varía con el contexto. Así, lo que en un corpus de oraciones se aísla como morfema no tiene por qué presentar el mismo tratamiento en otro corpus diferente.

Preferimos operar con este tipo de información variable obtenida del análisis del objeto a operar con información 100% fiable referida al 100% de los textos de un idioma contenida en cómodas bases de datos. Entendemos que lo realmente fiable y cómodo es el uso de reglas generales que no necesitan una adaptación diferente para cada idioma. Apliquemos nuestro programa a un ejemplo del inglés:

(8) aboxcontainsaboxthatcontainsabo...

Primera fase del análisis:

(8) abox / contains abox that / contains abox that / contains abox that / ...

Segunda fase del análisis:

(8) **the-box** / contains / **the-box** / that / contains / ...
(9) **my-box** / contains / **my-box** / that / contains / ...
(10) **a-box** / contains / **a-box** / that / contains / ...
(11) **abeautiful-box** / contains / **abeautiful-box** / that / contains / ...

Contexto γ

Obsérvese que las divisiones realizadas en la oración (8) son correctas.