Un sistema de análisis morfológico por ordenador

M.A.Martí
Intersoftware
Barcelona

En Intersoftware SA está en desarrollo desde 1977 un sistema de indexación automática consistente en el tratamiento de textos de manera que después de sucesivos análisis se llegue a una representación de su significado y posterior clasificación en un índice temático.

Para iniciar la investigación se entró una muestra de siete números del periódico "EL País" de los que se omitió todo lo que no fueran noticias de agencia.

Un primer estudio de esta muestra puso de manifiesto que el lenguaje de los textos periodísticos está sujeto a un proceso de renovación e incremento importante. Linguísticamente esto se traduce en la creación de nuevas palabras mediante la habilitación o la derivación principalmente.

También se observó que el significado asociado a los términos estaba sujeto con frecuencia a desplazamientos que lo llevaban lejos de su significado original.

La aparición de nuevas palabras, que en muchos casos no se encuentran en los diccionarios y ,por otra parte, la necesidad de "reconocer" el texto en su totalidad planteó la necesidad de diseñar una herramienta que resolviera el problema de la identificaión de formas, su análisis y la asociación de información a cada una de ellas. Se desechó desde un principio la entrada directa de palabras tanto por el volumen de lenguaje con el que tratábamos como por la incomodidad que posteriormente representaria el mantenimiento de este diccionario. Se decidió finalmente la construcción de un analizador morfológico (AM) que:

. evitara la entrada de formas una a una;
. facilitara el mantenimiento del corpus;
. permitiera asociar información a las unidades del análisis.

El AM de Intersoftware está basado en un **autómata markoviano ampliado con condiciones.** Se trata de un análisis izquierda-derecha que da como resultado todas las interpretaciones posibles para cada forma. El análisis se hace sobre la forma gráfica de las palabras.

La filosofía del sistema consiste en dar la combinatoria correcta de las raíces con los sufijos y morfemas de flexión. Para ello se ha de proceder a la segmentación de las formas y a dar la combinatoria mediante las reglas del autómata.

La información lingüística se expresa en forma de **atributos** del tipo : CAT=N, GEN=MASC, PERS=1, etc. que se asocian a las **unidades del diccionario (UD)** (raices, sufijos y morfemas de flexión) o bien a los **modelos.**

Los modelos son generalizaciones sobre el comportamiento

morfológico. Las UD están agrupadas en modelos de comportamiento morfológico homogéneo. Distinguimos dos tipos de modelos: modelos de raíces y modelos de sufijos y morfemas flexivos, es decir modelos que pueden aparecer en reglas que tienen el estado inicial START como estado de partida (modelos de raíces) y los que no (modelos de sufijos y flexión).

Sea <V1> el modelo de las raíces de la primera conjugación, <IP1> el modelo de morfemas de flexión del presente de indicativo de esta conjugación y <SDOR> el modelo del sufijo derivativo -ador. Las UD estarán distribuidas en cada uno de ellos de la siguiente manera:

<V1>	<IP1>	<SDOR>
am-	-o	-ador
salt-	-as	
mir-	-a	
par-	-amos	
lav-	-aís	
etc.		

fig. 1

No existe ningún tipo de condicionamiento en la definición de los modelos. El lingüista debe decidir el criterio según el cual agrupará las raíces en modelos: puede hacerlo en función de su comportamiento derivacional, de su paradigma de flexión, etc. En nuestro caso, como el paradigma derivacional es muy libre y nuestro objetivo era dar información morfológica, seguimos el criterio de agrupar las raíces según su paradigma de flexión. La información referente al paradigma de flexión va asociada al modelo ya que es común a todas las raíces mientras que la información sobre los sufijos que aceptan las raíces va asociada a éstas una a una.

```
<V1>
TV=V1
am-
salt-  SD=DOR
mir-   SD=DOR
par-
etc.
```

fig.2

En los modelos de flexión se opera de manera semejante:

```
<IP1>
T=PRES, MOD=IND
-o    PERS=1, NUM=SG
-as   PERS=2, NUM=SG
etc.
```

fig.3

Los atributos de tiempo y modo están en el modelo mientras que los de número y persona están asociados a los morfemas (la organización podía haber sido otra: modelos por personas con la información sobre modo y tiempo morfema a morfema, p.e.)

A lo largo del análisis se va recogiendo toda la información asociada al modelo o a las UD, de manera que cuando se consigue un análisis de una forma , ésta lleva asociados todos los atributos que se han encontrado:

"amo" TV=V1, PERS=1, NUM=SG, T=PRES, MOD=IND

Las reglas del autómata no operan con las UD sino con modelos y en general son de la forma:

ESTADO1-------<MODELO>-------->ESTADO2

fig.4

que si lo aplicamos a los modelos que acabamos de definir, tenemos (1):

R 001 START-------- <V1>---------> RA

R 002 RA ---------- <IP1>--------> RV

R 003 RA ---------- <SDOR>-------> RNA

fig.5

Así, la regla 1 tiene como estado de partida START y como estado de llegada RA (reconoce raíces); la regla 2 tiene RA como estado de partida y RV como estado de llegada (reconoce el verbo) ; la regla 3 tiene igualmente RA como estado de partida y RNA (reconoce nombres y adjetivos) como estado de llegada etc.

Según el analizador de la fig.5 obtendríamos como análisis correctos las formas "amador" y "lavador" cuando, según se desprende de la información asociada a las raíces, no hemos considerado que las UD "am-" y "lav-", deban combinarse con "-ador". Para ello es necesario poner condiciones a las reglas de manera que las formas como "amador" y "lavador" no sean aceptadas. El analizador de la fig.5 se ha de modificar según vemos a continuación:

R 001 START------ <V1>------>RA

R 002 RA -------- <IP1>----->RV condición: TV=V1

R 003 RA--------- <SDOR>---->RNA condición: SD=DOR

fig.6

---

(1) Para más detalles sobre el analizador: Ma.A.Martí "Un sistema d'anàlisi morfològica per ordinador" en Actas del I Congreso de lenguajes Naturales y Lenguajes Formales, Barcelona 1985.

El analizador no exige ningún tipo determinado de segmentación interna de las palabras, es el lingüista quien debe establecer los criterios en función de los objetivos del análisis . En nuestro caso optamos por reducir al máximo el número de entradas al diccionario siempre que no implicara complicar mucho la estructura del analizador. El verbo "poner" se podría segmentar, entre otras, de las siguientes maneras:

```
1). pong-  -o,..., -a, -as,..
2). pon-   -ía,..., -er,...
3). pus-   -e,..., -iera,..., -iese     fig.7
4). pondr- -é, ..., -ía
5). puest- -o, ...,-as.
```

o bien:

```
1). pon-  -g-   -o, ...
          -dr-  -é, ..., -ía
          -es,..., -en                  fig.8
2). pus-  -e, ..., -iera, ..., -iese
3). puest- -o, ..., -as.
```

En la fig.7 tenemos cinco raíces, pero el autómata sólo recorrerá dos reglas para analizar cada forma. En la fig.8, el número de raíces se reduce a tres, por lo tanto el diccionario está más comprimido, pero en el caso de "pongo" , "pondré" etc. el analizador tendrá que recorrer tres reglas para el análisis. En este último caso se puede objetar que se reduce el número de raíces pero aparecen dos nuevas UD: -g- y -dr- que nos dan un total de cinco entradas para formar las raíces.

Esta última opción la hemos tomado siempre que esos segmentos intermedios (-g-, -dr-) tienen un rendimiento distribucional mínimo, de manera que aunque representan dos UD nuevas por otra parte ahorran muchas entradas en el diccionario (todos los verbos que hacen el futuro y condicional con -dr- y el subjuntivo etc. con -g-): compondr-, compong-, vendr-, veng-, etc.

Además de la información morfológica las UD pueden llevar información referente a su significado léxico. El nivel de tratamiento del lenguaje que permite este analizador no hace aconsejable tratar la información referente al significado asociado a la raíz. En lingüística computacional se dispone de otros medios más adecuados (p.e. las redes semánticas) para resolver el problema. Lo que sí permite este nivel de análisis es tratar la información léxica asociada a los sufijos (1). Veámoslo con un ejemplo.

El sufijo -ador asociado a una raíz verbal suele formar nombres (y adjetivos) que designan el agente de la acción expresada por el verbo. En algunos casos (p.e.:"pasador", "colgador", etc.) da

---

(1) Es intereseante el planteamiento de esta cuestión en N.CERCONE "Morphological Analysis and Lexicon Design for Natural Language Processing", en Computers and the Humanities, vol.11, pp.235-258, 1978

lugar a un nombre de objeto y , en otros, a un nombre de lugar (p.e.: mirador, mostrador, etc.). Si queremos que la información asociada al sufijo se ponga de manifiesto en el análisis, en primer lugar tenemos que distinguir tres sufijos -ador distintos:

```
<ADOR1>          <ADOR2>          <ADOR3>
 IL= LUG          IL= OBJ          IL= AGEN
 -ador            -ador            -ador
```

fig. 9

(Crearemos modelos semejantes para las variantes -edor, -idor). Utilizamos el atributo IL para incorporar la información léxica, que en este caso puede tener los valores LUG (lugar), OBJ (objeto) y AGEN (agente, tanto nombre como adjetivo). Las raíces verbales deberán llevar un atributo referente a este sufijo para cumplir la condición de la regla correspondiente:

```
<V1>                          <V2>
am-
pas-   SN2=ADOR
mir-   SN1=ADOR
par-   SN1=ADOR
lav-
etc.                          com- SN1=EDOR, SN3=EDOR
```

fig.10

Para su análisis definimos la siguiente estructura del autómata:

```
R 001 START-------<V1>--------RA

R 002 START-------<V2>--------RA

R 003 RA----------<ADOR1>-----RNA cond.: SN1=ADOR
        analiza: "mirador", "parador", etc.

R 004 RA----------<ADOR2>-----RNA cond.: SN2=ADOR
        analiza: "pasador", etc.

R 005 RA----------<ADOR3>-----RNA cond.: SN3=ADOR
        analiza: "trabajador", etc.

R 006 RA----------<EDOR1>-----RNA cond.: SN1=EDOR
        analiza: "comedor1", etc

R 007 RA----------<EDOR3>-----RNA cond.: SN3=EDOR
        analiza: "comedor3", etc.
```

fig.11

Si queremos generalizar algún sufijo para todas las raíces de un mismo modelo basta con dar el atributo al modelo en lugar de darlo a las raíces una por una. Es una forma sencilla de preveer el análisis de formas posibles aunque muchas de ellas no se lleguen a usar nunca. Hemos limitado el uso de este recurso a los

sufijos con rendimiento distribucional elevado(p.e. -able para los verbos, -mente para adjetivos).

Otro tema interesante que plantea el análisis morfológico automatizado es el tratamiento de los homógrafos (tanto si proceden de una homonimia como de una polisemia). Si queremos dar información léxica a las palabras sometidas a análisis puede que lleguemos a tener que diferenciar la raíz "atrac-" que admite el derivado "-ador" ("Los atracadores retuvieron a los rehenes durante toda la mañana") de la raíz "atrac-" que no lo admite ("El barco atracó en la zona norte del puerto" e incluso "Se atraca de comida siempre que puede").

En estos casos los diccionarios semánticos (1) distinguen tres entradas distintas ya que cada una tiene una definición propia que difiere de las otras. En nuestro caso hemos optado por dar una sola entrada en el diccionario de raíces a partir de la cual se pueden formar todos los derivados correspondientes a cada una de las tres acepciones. Como el analizador morfológico es modular le puede ser suministrada esta información desde otros programas.

Actualmente está en proceso de implementación el tratamiento de las lexías ("máquina de escribir", "caña de pescar") y el de las perífrasis verbales a partir de información asociada a las raíces.

---

(1) I. Mel'cuk "Un nouveau type de dictionnaire: le dictionnaire explicatif et combinatoire du français contemporain". Cahiers de Lexicologie, v. 38, 1981, no.1

INSCRIPCION

SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

D.....

Dirección Postal
Calle:.....
Código postal:.....Ciudad.....
Teléfono.....

Centro de trabajo.....
Area de investigación o interés.....

Solicita su inscripción como socio de la S.E.P.L.N., para lo cual adjunta talón nominativo por valor de 1500 pst, correspondiente a la cuota del presente año.

Fecha.....Firma

ENVIAR A:

Jorge Garcia Sanz
Secretaria SEPLN
Clavel 12, bajo 3
Rivas-vaciamadrid
Madrid