Extracción del sintagma verbal núclea y resolución de ambigüedades en la asignación categorial

Gabriel G. Bès, Zulema Solana

Gabriel G.Bès

Zulema Solana

EXTRACCIÓN DEL SINTAGMA VERBAL NÚCLEO Y RESOLUCIÓN DE AMBIGÜEDADES EN LA ASIGNACIÓN CATEGORIAL

ABSTRACT

En el presente trabajo vamos a referirnos a la ambigüedad de la asignación categorial de las ocurrencias lingüísticas que son producto de la segmentación de un texto. La resolución de ambigüedad es una exigencia crucial en el análisis automático de textos.

Este problema será tratado en paralelo con la extracción de sintagmas núcleos, en particular el sintagma núcleo verbal, etapa indispensable en el análisis automático de la oración.

Recurriremos a dos útiles informáticos SMORPH (segmentación y morfología) y MPS (módulo post-smorf) integrados en una arquitectura común.

Presentaremos los resultados obtenidos a partir de un sondeo realizado sobre tres textos periodísticos con un total de 1200 palabras.

SUMARIO

0. Introducción

En el presente trabajo trataremos la extracción de sintagmas núcleos, en particular el sintagma núcleo verbal, etapa indispensable en el análisis automático de la oración en paralelo con el problema de la ambigüedad de la asignación categorial de las ocurrencias lingüísticas que son producto de la segmentación de un texto. La resolución de ambigüedad es una exigencia crucial en el análisis automático de textos.

Cuando hablamos de sintagma verbal núcleo nos estamos refiriendo al segmento del sintagma que comienza con el inicio de éste y finaliza con el núcleo, por ejemplo:

vio
lo vio
no lo vio

ha afirmado
lo ha afirmado
no lo ha afirmado

Los sintagmas núcleos poseen propiedades de linearidad que restringen su combinatoria interna. Conociendo las categorías morfo-sintácticas de las que están compuestos, es generalmente posible determinar su comienzo, su final y su núcleo.

El análisis en sintagmas núcleos permite reducir significativamente la ambigüedad de la categorización morfo-sintáctica al concatenar las expresiones internas. En

los ha comprado o en *comprarlos*
los no puede ser artículo

Vamos a recurrir a las posibilidades de este análisis para lograr desambigüizar la mayor parte de los casos que presentaremos.Consideraremos los tipos de ambigüedad en el sintagma verbal núcleo, en especial la ambigüedad del verbo en tercera persona, y su posible resolución.

Describiremos y haremos uso de dos útiles de tratamiento informático: SMORPH (segmentación y morfología) y MPS (módulo post-morf) dos módulos que pertenecen a una arquitectura común

Presentaremos los resultados obtenidos a partir de un sondeo realizado sobre tres textos periodísticos con un total de 1200 palabras. Propondremos la modelización al respecto y discutiremos los resultados obtenidos

1. Tipos de ambigüedad en el sintagma verbal núcleo

1.1. Verbos ambiguos en la tercera del singular.

Se observa que a marcas morfológicas menos específicas hay mayor posibilidad de ambigüedad. Un verbo terminado en –amos, por ejemplo, es muy difícil que pueda ser homónimo de otra palabra de la misma forma y otra asignación categorial, en cambio un verbo que termina en –a entra en ambigüedad con

un sustantivo en –a (Juan ama /el ama de llaves)
o un adverbio en -a. (Juan cerca su casa/ Queda cerca)

En este sentido, los casos más frecuentes son las terceras personas singulares de los presentes de indicativo y subjuntivo.

Por ejemplo:

recuerdo (v)
recuerdo (n)

vela (v)
vela (n)
la vela (cl + v)
la vela (det + n)

cuenta (v)
cuenta (n)

la cuenta (cl + v)
la cuenta (det + n)

Se trata siempre de verbos en -ar que admiten clítico acusativo y pasiva (transitivos en términos tradicionales), que están en presente de indicativo. Se observa que los sustantivos están formados con sufijo -a y el verbo también

Además existen ambigüedades en verbos en -er que admiten clítico acusativo y pasiva, y están en presente de subjuntivo. Se observa que los sustantivos están formados con sufijo -a y el verbo también.

coma (v)
coma (n)

la coma (cl + v)
la coma (det + n)

beba (v)
beba (n)
la beba (cl + v)
la beba (det + n)

NO HAY AMBIGÜEDAD cuando el n se forma con sufijos especializados

declara (v)
declaración (n)

pinta(v)
pintor(n)
pintura(n)

teme(v)
temor (n)

1.2. Ambigüedad de los clíticos antepuestos (proclíticos)

1.2.1. Clíticos simples

me, nos, os, le y les no son ambiguos

te y *se* son ambiguos si no se tiene en cuenta el acento, no lo son si se toma en consideración el acento:

-tomo té (n)
-te digo (cl)

-sé poco de eso (v)
-se cayó (cl)

lo clítico y *lo* determinante, *los* clítico y *los* determinante, *la* clítico y *la* determinante, *las* clítico y *las* determinante son ambiguos

-lo vio (cl)
-lo bueno (det)

-los vio (cl)
-los niños (det)

-la vio (cl)
-la niña (det)

-las vio (cl)
-las niñas (det)

1.2.2. Combinaciones de clíticos

En la oralidad podrían ser ambiguos por ejemplo :

-te (cl) me (cl)
-teme (v)

pero en la escritura el espacio en el medio impide la ambigüedad. Igualmente con la pronunciación rioplatense serían ambiguos :

-se (cl) los (cl)
-celos (n)

1.3. Auxiliares

1.3.1. Auxiliar haber

Haya o *había* pueden tener distinta asignación categorial

-haya escrito (aux + participio)
-planté un haya (det +n)
-había dicho (aux + participio)
-había tormenta (v + n)

1.3.2. Auxiliar ser

-son amados (aux)
-el son de la guitarra (n)
-son buenos (v)

-era(s) amado (aux)
-era bueno (v)
-la era paleozoica (n)

2. Resolución posible de ambigüedades

2.1. El verbo

PRIMER CASO

La ambigüedad se resuelve al anteponerle el determinante ‘el’ o el clítico ‘lo’ porque el sustantivo es masculino y el determinante masculino que concuerda con él no es ambiguo respecto de ningún clítico.

recuerdo (v)
recuerdo (n)

lo recuerdo (cl + v)

el recuerdo (art + n)

En estos ejemplos la construcción del sintagma nominal núcleo o del sintagma verbal núcleo desambigüiza

SEGUNDO CASO

La ambigüedad no se resuelve al anteponerle determinante o clítico porque el sustantivo es femenino y, en consecuencia determinante y clítico también pueden ser ambiguos cuando son femeninos.

A continuación presentamos dos ejemplos cuyo esquema se repite frecuentemente.

vela (v)
vela (n)
la vela (cl + v)
la vela (det + n)

Se desambigüiza cuando al clítico ambiguo le antecede otro clítico o una negación y al n ambiguo le sigue un adjetivo (o un verbo)

-no la(cl) vela(v) según sus ritos
- la (det) vela(n) encendida(adj)
-la(det) vela(n) parpadea(v)

cuenta (v)
cuenta (n)
la cuenta (cl + v)
la cuenta (det + n)
-no la(cl) cuenta(v) con detalle
-la(det) cuenta(n) de cristal
-la(det) cuenta(n) de una cifra

Si la negación antecede al clítico ambiguo quita la ambigüedad y si al sustantivo ambiguo le sigue un complemento con ‘de’ desaparece la ambigüedad

TERCER CASO

En este caso (menos productivo), se trata de verbos en -er que admiten clítico acusativo y pasiva, y están en presente de subjuntivo. Se observa que los sustantivos están formados con sufijo -a y el verbo también.

coma (v)
coma (n)

la coma (cl + v)
la coma (det + n)

-no(neg) la(cl) coma(v)
-la(det) coma(n) indicada(adj)
(-la(det) coma (n) señala (v))

beba (v)
beba (n)

la beba (cl + v)
la beba (det + n)

-no(neg) la(cl) beba(v)
-la(det) beba(n) pequeña (adj)
(-la(det) beba (n) llora (v))

2.2. Auxiliares

2.2.1. Auxiliar haber

La ambigüedad desaparece si se considera la palabra anterior o posterior o se antepone o pospone una palabra.

Por ejemplo: *haya* o *había* pueden tener distinta asignación categorial

-haya escrito (aux + participio)
-planté un haya (det +n)
-había dicho (aux + participio)
-había tormenta (v + n)

pero la ambigüedad desaparece dentro del sintagma núcleo sea verbal o nominal

2.2.2. Auxiliar ser

En este caso ocurre como en el anterior, la ambigüedad desaparece en el ámbito del sintagma núcleo

3. Útiles de tratamiento informático

3.1. El analizador morfosintáctico SMORPH (cf: Bès- Rodrigo 01,)

Smorph (*segmentación* y *morfología*), es el software especificado e implementado en el GRIL por Salah Aït-Mokhtar, trata en una sola etapa todo lo relativo a la *pre-sintaxis*. Esta herramienta totalmente declarativa recibe en entrada una secuencia de códigos Ascii y, da a la salida una secuencia de ocurrencias asociadas a un conjunto de pares <*etiqueta=valor*>.

Smorph es perfectamente declarativo. Se deben especificar en él cinco tipos de información:

i-Los rasgos <etiqueta; {v1, ..., vn}> que se quieren utilizar. Por ejemplo, que la etiqueta NUM (número) va a tener como valores posibles sg y pl(singular y plural).

ii-Las terminaciones que se utilizarán (por ej. la terminación *–es* para *trae<u>s</u>* y *rosale<u>s</u>*).

iii-Los códigos Ascii que se quieren utilizar como separadores.: el espacio (que tiene el número. 32), y otros clásicos como el salto de línea, además ’, ‘:’, y : ‘?’, ‘!’que por otra parte son códigos Ascii que se imprimen (no así el espacio en blanco).

iv-Los modelos, que mediante la concatenación de cadenas adyacentes, permiten generar ocurrencias complejas. En la flexión verbal se obtiene *amo, amas, ama ...* a partir de la concatenación del radical *am* con las terminaciones *o, as, a* y los objetos generados de este modo son reconocidos como ocurrencias del lema *amar* y se les asocian los rasgos que les corresponden (rasgos de persona, número, tiempo, modo etc). Las operaciones y los rasgos

se habrán declarado en el modelo correspondiente, que será el mismo para un subconjunto de entradas.

v-Las entradas, donde se declararán las formas individuales con indicación de los modelos en el caso de generación regular (es el caso de *amar, cantar, llorar etc.*) o con la indicación de los rasgos propios para las formas que no se pueden generar de manera regular. Por ej. la expresión *los*, que, en dos entradas diferentes, va a recibir los rasgos propios del artículo definido plural en una y del pronombre clítico de 3ª persona plural en otra.

Todo lo anterior se declara mediante una sintaxis clara y explicitada. Se trata de información totalmente declarativa, disociada de su tratamiento algorítmico. Quien desee tratar una nueva lengua no necesita introducirse en el código informático de Smorph. Lo único que debe hacer es especificar las declaraciones pertinentes de los cinco tipos de información antes mencionados y Smorph hace el resto.

Cuando una secuencia no es ambigua, a un lema [1] se asocia una categoría pero lo que ocurre frecuentemente es que si a cada ocurrencia lingüística la tomamos aisladamente, que es lo que hace SMORPH (y, por otra parte, cualquier "etiquetador", o cualquier analizador morfológico), puede resultar ambigua. Así a la ocurrencia lingüística "la" se le asignarán dos etiquetas (clítico y artículo) y a "ayuda" las de nombre y verbo, lo que implica dos representaciones semánticas.

En este trabajo vamos a considerar las situaciones de ambigüedad que se producen con los elementos que constituyen el sintagma verbal núcleo.(cf.Rodrigo 01 y Bès-Rodrigo 01- 04) y se propondrán para la desambigüización reglas de recomposición de MPS

3.2. MÓDULO MPS

Cuando trabajamos con textos reales nos encontramos con expresiones que no se adaptan a una sintaxis estándar. Entre estas expresiones podemos contabilizar fechas, cantidades, todo lo que concierne a la sufijación y prefijación, incluido el tratamiento de clíticos, y las contracciones en lo que en algún momento se llamó morfemas "complejos". Si se busca normalizar la entrada de la sintaxis para tener expresiones que satisfagan las mismas relaciones. parece útil tratarlos en una etapa anterior al del resto de la sintaxis general de la frase.

El módulo post-smorph, (MPS [2]), analiza estos microsistemas. MPS recibe en entrada una salida Smorph (en formato Prolog) y va a dar en salida otro formato según el analizador que se vaya a utilizar; MPS podrá modificar las estructuras de datos recibidos en la entrada. MPS ejecuta dos funciones principales: la Recomposición y la Correspondencia; la Recomposición a su vez puede ser de dos tipos diferentes, el Reagrupamiento y la División.

---

[1] Bès, Rodrigo 01 4.2. nota 27: " (lema) es una etiqueta arbitraria elegida para caracterizar un conjunto de raíces en relación con un conjunto de anteposiciones o terminaciones. La tradición hace que se elija la forma verbal del infinitivo para representar los lemas verbales, la forma singular de los nombres para los lemas nominales, etc."

[2] implantado por Faiza Abbaci(cf. Abbaci 99)

MPS es una herramienta declarativa, en donde mediante reglas se pueden expresar los valores de entrada (sobre dos o más estructuras de datos de la salida Smorph) y los valores de salida sobre la estructura reagrupada.

Las reglas que se declaren con la función de división van a provocar el efecto inverso. Son útiles para tratar las contracciones (por ej. una ocurrencia de *del* en español), afín de obtener en la salida una secuencia de entidades que sea análoga a las que Smorph asigna a las ocurrencias no contraídas en una cadena.

Las reglas que se declaren con la función de correspondencia van a operar sobre una sola estructura de datos a la salida de Smorph y van a poder modificarla en otra estructura de datos. Estas reglas permiten formular en Smorph descripciones básicas, generales, y adaptarlas después a la exigencia de cada analizador o de cada aplicación, o enriquecerlas con nuevos pares de <etiqueta=valor>.

3.3. REGLAS DE DESAMBIGÜIZACIÓN

3.3.1. Tomaremos en consideración uno de los casos planteados

A continuación tenemos, analizado por SMORPH a:

-no la vela

podrá verse que presenta los dos análisis posibles para ‘la’(artículo y clítico) y para ‘vela’(verbo y sustantivo).

```
'no'.
[ 'no', 'TADV','neg'].
'la'.
[ 'el', 'EMS','art'].
[ 'lo', 'TPRON','cl'].
'vela'.
[ 'vela', 'EMS','nom', 'NUM','sg'].
[ 'velar', 'EMS','v', 'MVERB','fl', 'NUM','sg'].
'.'.
[ '.', mi].
```

Si esta expresión analizada es el input de MPS y se le aplica la siguiente regla de recomposición.

```
%neg+clit+v da svn%
S0 [L0, 'TADV', 'neg'] S1 [L1, 'TPRON', 'cl'] S2 [L2, 'EMS', 'v']
-->S0+S1+S2 [L0+L1+L2, 'EMS', 'svn'].
```

Tenemos:

```
'no la vela'.
[ 'no lo velar', 'EMS', 'svn' ].

'.'.
[ '.' , mi ].
%fin de la phrase numero : 1
```

Es decir, hemos desambigüizado la expresión

Si expresamos las ocurrencias por variables tenemos las siguientes reglas:

Regla1a
X[neg] Y[cl] Z[va] → X+Y+Z[SVn]

Regla 1b
X[neg] Y[cl] Z[va] → X [neg] Y[cl] Z[v]

De la misma manera podemos proceder con 'la vela encendida' y tendremos:

Regla 2ª
X[art] Y[na] Z [adj] → X + Y + Z [SNn]

Regla2b
X[art] Y[na] Z [adj] → X[art] Y[n] Z [adj]

4. Modelización propuesta

La modelización propuesta sigue una línea de trabajo ya propuesta para tratar problemas análogos del francés (cf. Bès et al, 2004). En francés, como en español, por ejemplo, formas proclíticas son ambiguas con respecto a artículos (es el caso del francés *la*) y formas verbales son ambiguas con respecto al nombre (es el caso del francés *juge*).

Contrariamente a técnicas basadas en probabilidades de transición que exigen la utilización de grandes corpus etiqueteados, se trata de rentabilizar al máximo una modelización del sistema lingüístico, en el marco de la exigencia general siguinte: obtener el máximo de resultados con el mínimo de recursos.

Nuestros recursos han sido incorporados a Smorph y a MPS.

En Smorph se dispone como punto de partida de una base de entradas léxicas con todas las expresiones que pueden ocurrir en los sintagmas verbales núcleos: proclíticos, formas verbales flexivas y participiales, negación, auxiliares. Las formas ambiguas no han sido declaradas mediante entradas múltiples diferentes, asociadas a rasgos diferentes, sino mediante una entrada única asociada a rasgos que indican una ambigüedad potencial. Así, por ejemplo, la expresión *la* no está representada mediante dos entradas diferentes y diferenciadas mediante los rasgos de cada una de ellas, sino que está representada en las entrads léxicas por una forma única asociada a un rasgo (en la ocurrencia *ambcl*) que la caracteriza como una forma proclítica potencialmente ambigua. *Mutatis mutandis*, las expresiones *ayuda* o *trabajo* están representadas por una entrada única y caracterizadas, mediante el rasgo *ambv*, como formas verbales potencialmente ambiguas con respecto a nombres.

En MPS se han declarado reglas estructurales y reglas heurísticas. Las primeras están motivadas por una modelización lingüística simple del español y están destinadas a resolver contextualmente las ambigüedades de categorización, al mismo tiempo que, la mayor parte entre ellas, especifican los sintagmas verbales núcleos. Las segundas están motivadas por consideraciones heurísticas simples y se aplican a los casos límites que quedan fuera del alcance de las reglas estructurales.

Hay dos grandes tipos de reglas estructurales: aquellas que especifican que una expresión – ambigua o no – pertenece a un sintagma verbal núcleo, ya sea como forma clítica o como verbo, y aquellas que especifican que una expresión potencialmente ambigua no pertenece a un sintagma verbal núcleo.

En los ejemplos que siguen, X, Y, Z son variables sobre las ocurrencias y los rasgos se expresan de manera muy simplificada).

Ejemplo de regla estructural para especificar sintagmas verbales núcleos

Si en la entrada a analizar se tiene:

X[clitico] Y[haber] Z[verbo en forma participial]

entonces, formar el sintagma verbal núcleo *X+Y+Z.*

Esta regla va a formar los sintagmas verbales núceos de, p.ej. *Ella (la ha visto), Nosotros (los hemos visto).*

Ejemplo de regla estructural para especificar la no pertenencia de una forma potencialmente ambigua a un sintagma verbal núcleo

Si en la entrada a analizar se tiene:

X[articulo indef.] Y[ambv]

entonces, formar el sintagma nominal núcleo *X+Y*, lo que va excluir *Y[ambv]* como forma verbal.

Esta regla va a formar los sintagmas nominales núcleos de, p.ej., *una ayuda*, *un trabajo.*

La reglas heurísticamente motivadas son aquellas que se aplican a expresiones no analizables, en la modelización adoptada, mediante las reglas estructurales. Un ejemplo es el tratamiento de expresiones como *trabajo* cuando ocurren fuera de los sintagmas nominales núcleos analizables por las reglas estructurales (por ejemplo en la expresión *con constancia y trabajo*).

5. Resultados obtenidos

La modelización propuesta y declarada en el sistema Smorph-MPS fue aplicada a un sondeo efectuado sobre textos periodísticos. El corpus utilizado para el sondeo fueron artículos tomados de los diarios *Clarín*, *La Nación* y *Página/12*, de ahora en más identificados por *T1*, *T2* y *T3* respectivamente.

La tabla siguiente da el número de palabras y de ocurrencias de sintagmas verbales núcleos - N(*svn*) - en los textos analizados

	N(palabras)	N(svn)
T1	396	51
T2	248	25
T3	567	52
Total	1.211	128

Los resultados obtenidos se resumen en la tabla siguiente, en donde :

N1 : N(svn correctamente analizados)

N2 : N(svn analizados)

Precisión : N1/N2 x 100

Cobertura: N1/N(svn) x 100

	N1	N2	Precisión	Coberura
T1	50	50	100%	98,0%
T2	25	25	100%	100%
T3	51	51	100%	98,0%
Total	126	126	100%	98,4%

Observemos que los dos errores (sobre 128 posibles) provienen de reglas fundadas heurísticamente y sólo afectan la cobertura.

6. REFERENCIAS

Bès G.G.,Lamadon L.,Trouilleux F.(2004) « Verbal chunk extraction in French using limited resources » arXiv :cs.CL/0408060 v1

Bès G. G. (2002): "La linguistique entre science et ingénierie". En *TAL* Vol. 41 n. 3.

Bès G.G.,Solana Z.(2004)«Los clíticos del español. Tipos de verbos y duplicación » comunicación presentada ante al CONGRESO CÁTEDRA UNESCO, Buenos Aires

Rodrigo Mateos, J. L., Bès G.G. (2003): "Análisis e implementación de clíticos en una herramienta declarativa de tratamiento automático de corpus". Comunicación aceptada para el *VI Congreso de Lingüística General*, Santiago de Compostela, España, 2