Lluís de Yzaguirre Maura
Anna Matamala Ripoll
Carme Bach Martorell
Núria Castillo Igea
Eugènia Ustrell Peñafiel

Institut de Lingüística Aplicada
Universitat Pompeu Fabra
de_yza@upf.es
www.iula.upf.es

AMBILIC, EL DESAMBIGUADOR LINGÜÍSTICO DEL CORPUS DEL IULA (UPF)

(XVII Congreso AESLA, Lingüística de Corpus y Computacional)

ABSTRACT

AMBILIC es un programa de desambiguación de base lexicomorfológica para el catalán y para el castellano que se utiliza en el tratamiento del corpus técnico del Institut Universitari de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra (UPF), después de las fases de marcaje estructural, preproceso y análisis morfológico. En esta comunicación presentamos el funcionamiento del AMBILIC y los resultados que ofrece. Nos fijamos especialmente en el tipo de reglas que aplica, diferenciando entre las contextuales y las acontextuales. También explicamos brevemente el tratamiento de casos especiales --como las ambigüedades segmentales, las mayúsculas o las personas verbales—y, finalmente, indicamos mejoras relacionadas con la subespecificación y otras de base sintáctica, temática, dialectal y terminológica.

1.- Introducción

AMBILIC es un programa de desambiguación de base lexicomorfológica para el castellano y para el catalán que se utiliza en la cadena de tratamiento del corpus técnico[1] del Institut de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra. Se aplica a ficheros analizados morfológicamente con el PALIC[2], con la finalidad de eliminar el máximo número de ambigüedades con reglas únicamente lingüísticas. Los resultados demuestran que AMBILIC reduce las ambigüedades de un 170% (porcentaje inicial, similar, v.g., al del francés[3]) a un 110% aproximadamente. Las que quedan después de

[1] Véase Bach, C. *et al.* (1997).
[2] véase De Yzaguirre (2000a)
[3] Véase Tzoukermann *et al.* (1996)

aplicar AMBILIC se resuelven mediante un desambiguador estocástico[4].

En esta comunicación presentaremos de modo general el funcionamiento interno del programa AMBILIC y los tipo de reglas que utiliza, así como posibles mejoras de cara al futuro.

2.- Funcionamiento de AMBILIC y resultados

AMBILIC procesa textos que han sido analizados morfológicamente y que, por lo tanto, contienen el/los lema(s) y la(s) categoría(s) de cada forma. Las categorías están codificadas mediante unas etiquetas establecidas para el corpus del IULA siguiendo las directrices del proyecto EAGLES[5]. Además, en los textos del corpus del IULA, los documentos están marcados estructuralmente[6] y han pasado por un preproceso en el que se han detectado y etiquetado algunas unidades particulares, como abreviaturas, números, locuciones, nombres propios, etcétera. Después de aplicar el programa a estos textos, el resultado es un documento donde están desactivados los lemas y las categorías que no son pertinentes según el contexto. El programa lo discrimina gráficamente indicando con minúscula la etiqueta que no es posible, mientras el resto de opciones siguen representadas con mayúscula.

En el marco del proyecto Corpus, distinguimos entre ambigüedad y subespecificación. Si una palabra tiene dos interpretaciones, pero ambas corresponden al mismo lema, le atribuímos una etiqueta más genérica, que llamamos subespecificada. Así, el sustantivo "viernes" es etiquetado N5M6, que significa "nombre común masculino de género subespecificado", o sea indistintamente singular o plural. Lo mismo hacemos con adjetivos como "bestial" o con formas verbales como "lleva" o "diría". Sólo consideramos ambiguas y sólo nos proponemos desambiguar aquellas formas con varias interpretaciones atribuíbles a lemas distintos.

AMBILIC resuelve las ambigüedades utilizando gramáticas locales[7] en forma de bancos de reglas del catalán o del castellano. De acuerdo con la posición que AMBILIC ocupa en la cadena de procesamiento del Corpus del IULA, sus reglas han sido formuladas restrictivamente; siempre es preferible que no actúen si el contexto no está muy bien definido y que las ambigüedades se resuelvan en el paso siguiente --la desambiguación por métodos estocásticos.

En lo que concierne a los resultados, en un estudio preliminar con un corpus de cien mil palabras, el margen de error del PALIC y del AMBILIC combinados ha sido de un 3%. Ambos se han aplicado a un total de más de cinco millones de palabras con un funcionamiento satisfactorio.

---

4 Véase Armstrong, S. *et al.* "Building a language model for POS tagging" (http://issco-www.unige.ch/tools).
5 Véase Morel, J. *et al.* (1997)
6 Véase Vivaldi, J. *et al.* (1996)
7 véase Silberztein (1993)

3.- Tipo de reglas

El AMBILIC opera con paquetes de reglas que se agrupan cuando tienen rasgos similares: esto facilita la ejecución del programa y también permite que el usuario pueda activar o desactivar un determinado paquete teniendo en cuenta las características textuales del documento. Hay distintos criterios de agrupación de reglas: por un lado, se pueden reunir en un mismo paquete según sean de tipo gramatical, de tipo léxico o híbridas. Por otro lado, se pueden agrupar en base a características dialectales (reglas dialectales) o temáticas (reglas temáticas). Un ejemplo de estas últimas serían las referentes a los nombres de notas o a los nombres de letras.

El programa AMBILIC asocia a cada rasgo de las etiquetas morfológicas un conjunto de descriptores de rasgos, de la forma siguiente:

A{MF}{SP}	/categoría, género, número
EF{MF6}{SP}	/categoría, clase, género, número

En el primer ejemplo, se nos indica que la categoría "A" puede tener género masculino o femenino y que el número puede ser singular o plural. En el segundo, observamos que la categoría "E" de la clase "F" puede tener género masculino, femenino o indefinido y que el número puede ser singular o plural. Estos descriptores basados en el etiquetario son los que permiten redactar reglas, además de los que permiten formular condiciones basadas en formas, lemas, número de lemas, posición en el contexto (inicial, final...)

Las reglas pueden tomar distintas decisiones al mismo tiempo y constan de dos partes: la condicional y la ejecutiva. La primera filtra si la regla se tiene que ejecutar o no, y la segunda sirve para eliminar un lema o unos lemas determinados o para eliminarlos todos menos uno. Por ejemplo:

```
* Regla 0001 -> "a"
0,lema=a
\
0,categoría=P
/
```

Para leer estas reglas, debemos tener en cuenta que, en cada línea, la cifra nos indica la distancia de la palabra a la que se refiere la condición o ejecución respecto a la primera del contexto que estamos estudiando. A continuación, se sitúa el descriptor de la variable lingüística (en este ejemplo, "lema" o “categoría”) seguido de un delimitador, que puede ser "=" (es) o "#" (no es). Al final de la línea aparece el valor que tiene que presentar la variable para actuar. La contrabarra "\" indica que se acaba la parte condicional y empieza la ejecutiva, y la barra "/" indica el final de la regla.

Los paquetes de reglas del programa AMBILIC se aplican a todas las unidades que se encuentran entre dos signos de puntuación siguiendo un orden secuencial, de izquierda a

derecha. Si cuando se llega a la posición final se ha modificado alguna palabra, se vuelven a aplicar todas las reglas del paquete y así sucesivamente hasta que no haya ninguna modificación[8]. El usuario determina qué paquetes se aplican, en qué orden e incluso cuáles se aplican más de una vez.

Las reglas pueden ser de dos tipos:

3.1.- reglas acontextuales

Son las que eliminan alguna interpretación sin tener en cuenta el contexto, como por ejemplo:

```
* Regla 0001 -> "a"
0,lema=a
\
0,categoría=P
/
```

La regla del ejemplo nos indica que, cuando encontramos un lema "*a*", se desambiguará como preposición. En consecuencia, se anula la posibilidad de que "*a*" se considere sustantivo.

Las reglas acontextuales se utilizan en casos como los nombres de notas o los nombres de letras, en vez de eliminar estas unidades del diccionario del analizador morfológico. El resultado en ambos casos es el mismo, pero este sistema permite que, en textos concretos, el usuario pueda desactivar estas reglas. Por ejemplo, podemos formular unas reglas acontextuales según las cuales las unidades "*a*", "*de*", "*e*", "*ca*", "*ele*", "*o*", "*erre*", "*ese*" o "*te*" no son sustantivos –y anulamos así la posibilidad de que correspondan a nombres de letras--, pero el usuario tiene la opción de desactivar estas reglas cuando desambigüe un texto de temática lingüística, en el que es probable que aparezcan como nombres de letras y, por lo tanto, como sustantivos.

3.2.- reglas contextuales

Tienen en cuenta el contexto a la hora de formular las condiciones. Veamos unos ejemplos

`* Regla 4004 -> "representan apenas"`
`0,tiempo=W`
 	
`1,mot=apenas`
`\`
`1,categoría=D`
`/`
	`Interpretación:`
	`Si la palabra observada es una forma verbal personal,`
	`seguida de la palabra “apenas”,`
	`entonces`
	`“apenas” es adverbio`

[8] En otras palabras, la condición de finalización del ciclo es el ciclo vacuo...

* Regla 4005 -> "el mejor de" 0,categoría=A 1,n_lemes=2 1,categoría=J 1,categoría=D 2,categoría=P \ 1,categoría=J /	Interpretación: Si la palabra observada es artículo, va seguida de palabra con dos lemas, o bien adjetivo o bien adverbio, seguido a su vez de preposición, entonces la siguiente es adjetivo
*Regla 4019 -> "la requerida" 0,n_lemes=2 0,categoría=A 0,categoría=R 1,n_lemes=1 1,modo=C \ 0,categoría=A /	Interpretación: Si la palabra observada tiene dos lemas, uno de ellos artículo y el otro pronombre y la palabra siguiente tiene un lema y es participio entonces la palabra observada es artículo
* Regla 4026 -> "cuyos intereses" 0,lema=cuyo 1,n_lemes=2 1,categoría=V 1,categoría=N \ 1,categoría=N /	Interpretación: Si el lema observado es “cuyo”, y la palabra siguiente tiene dos lemas, o bien verbo o bien nombre, entonces la palabra siguiente a la observada es nombre
* Regla 4030 ->art. seguido de nombre 0,categoría=A 1,n_lemes=1 1,categoría=X{NJE} \ 0,categoría=A /	Interpretación: Cualquier palabra que pueda ser artículo y vaya seguida de palabra unívoca cuya categoría sea nombre, adjetivo o especificador, entonces esa palabra es solamente artículo
* Regla 4033 -> pronombre#art? seguido de no verbo 0,n_lemes=2 0,categoría=R 0,categoría=A 1,categoría#V \ 0,categoría=A /	Interpretación: Si una palabra con dos interpretaciones una de ellas pronombre y la otra artículo va seguida de palabra que no es verbo entonces esa palabra es solamente artículo
* Regla 4043 "el acceptante" 0,categoría=A 1,n_lemes=2 1,categoría=X{NJ} 1,modo=G \ 0,categoría=A 1,categoría=X /	 Si un artículo va seguida de una palabra con dos lemas, uno de ellos nombre o adjetivo y el otro gerundio entonces el artículo sólo es artículo y la siguiente sólo es no gerundio

El formalismo de las reglas de AMBILIC permite expresar variables coincidentes (como X en la 4043 precedente) o que una determinada variable puede tener cualquier valor (tiempo=W indica que se trata de una forma verbal conjugada, lo que excluye las formas nominales de los verbos y el resto de categorías, como en la 4004). También permite

expresar una lista de lemas o de formas en una única condición; por ejemplo ante coincidencias concurrenciales del tipos “derechos humanos, individuales, colectivos, inenajenables, fundamentales” se puede proponer:

```
0,lema=derecho
0,género=X{SP}
1,lema={humano, individual, colectivo, inenajenable, fundamental}
1,género=X
\
0,categoría=N
1,categoría=J
/
```

En este caso, si las dos palabras son consecutivas y concuerdan en género, serán consideras respectivamente nombre y adjetivo.

3.1.- Casos especiales

En determinadas situaciones (por ejemplo, para entrenar el desambiguador estocástico) se necesitan materiales desambiguados al 100%, de modo manual. A menudo sucede que durante la desambiguación manual, posterior a la lematización y a la desambiguación automática, se descubren deficiencias que aconsejan repetir el proceso después de ser corregidas. A partir de esta desambiguación manual, AMBILIC puede generar bancos de reglas que sirven para reprocesar los documentos. Este conjunto de *reglas manuales* forman un paquete especial que se puede aplicar después de las reglas automáticas.

En cuanto a las *ambigüedades segmentales*, el programa permite que se formulen reglas específicas para resolverlas[9]**.** Nos referimos a ambigüedades como las que presenta la forma verbal "*verse*" (que se puede interpretar como forma del verbo "*versar*" o como forma del verbo "*ver*" más enclítico).

Además, el programa contempla dos casos especiales que no están formalizados como reglas sino que están incorporados como procesos. Sin embargo, el programa los reconoce como paquetes de reglas y es el usuario quien escoge si tienen que actuar o no. Nos referimos a las mayúsculas y a las personas verbales.

(a) las *mayúsculas*: si se detecta una palabra con mayúscula y una de las interpretaciones es un sustantivo, entonces se desambigua como sustantivo. Nos referimos a unidades como "*Juramento Hipocrático*", en la que "*juramento*" puede ser sustantivo o primera persona del presente de indicativo del verbo "*juramentar*". Si se activa el procedimiento de desambiguación por mayúsculas, el programa lo desambigua automáticamente como sustantivo. Otro ejemplo lo encontramos en la frase "*Hemos llegado a La Puebla*". Una vez más, el programa desambigua "*Puebla*" como sustantivo y no como tercera persona del presente de indicativo de "*poblar*" porque detecta las mayúsculas.

---

[9] véase De Yzaguirre (2000b)

(b) las *personas verbales*: a menudo sucede que, en un texto técnico redactado íntegramente en tercera persona, muchas ambigüedades tienen entre sus interpretaciones una forma verbal en primera o segunda persona. El procedimiento de desambiguación de personas verbales, en caso de ser activado por el usuario, analiza si en la totalidad del documento hay primeras o segundas personas verbales no ambiguas, en cuyo caso se inhibe de actuar; por el contrario, si las únicas primeras o segundas personas halladas corresponden siempre a formas ambiguas, como "*juramento*", "*base*", "*cierre*" o "*suplemento*", elimina la interpretación verbal, siempre y cuando el documento tenga una cierta extensión y una cierta proporción de formas verbales no ambiguas.

4.- Posibles mejoras del desambiguador

Hay muchas mejoras del programa que aún se tienen que desarrollar y que presentamos brevemente a continuación:

(a) Desambiguación de base sintáctica: se trata de interponer, entre el desambiguador morfológico y el estocástico, un parser sintáctico parcial[10] o *chunker*, al cual se le pasen tantas copias de una misma frase como combinaciones se puedan formar a partir de las ambigüedades aún presentes. El *chunker* nos devolverá resultados más o menos satisfactorios para cada una de las interpretaciones. Conservando los mejores resultados, una parte de las ambigüedades pueden ser eliminadas por cuanto todos los resultados preferidos comparten la misma interpretación de una determinada ambigüedad. De momento no es posible aplicar esta técnica porque no disponemos de un *chunker* suficientemente veloz para analizar los miles de combinaciones producidas por cada frase del corpus técnico del IULA. Sin embargo, dentro de poco tiempo las mejoras en los equipos usados y en los parsers harán rentable esta técnica.

(b) Desambiguación de base temática: se trata de implementar un procedimiento que permita detectar automáticamente si el tema de un documento impide aplicar un paquete de reglas concreto, por ejemplo los ya citados de notas musicales o nombres de letras.

(c) Mejoras de base dialectal: el programa tendría que poder detectar automáticamente aquellas variantes dialectales de un texto para las que existan paquetes de reglas. Así, por ejemplo, en un texto escrito en el Perú, "*lustrada*" tendría que figurar como participio del verbo "*lustrar*" y como sustantivo. En cambio, en un texto escrito en otra variedad, se podría desambiguar automáticamente como participio.

(d) Mejoras de base terminológica: para poder explotar a fondo la desambiguación de base temática, habría que desarrollar un programa que fuera capaz de reutilizar terminografías --especialmente en los términos sintagmáticos-- para generar reglas de desambiguación para un determinado ámbito terminológico. Por ejemplo, si está desambiguando un texto sobre informática y localiza la unidad "*base de datos*", tendría que poder desambiguar "*base*" como sustantivo y no como forma verbal.

---

[10] véase Hindle, D. (1994)

(e) Resolución de la subespecificación: como hemos visto, las reglas sólo eliminan interpretaciones de entre las atribuidas por el lematizador, pero no añaden ni modifican nada. Así, pues, queda por resolver la cuestión de las subespecificaciones. Nos referimos a casos como la forma verbal "*andaba*", en la que la persona es primera o tercera. Las condiciones contextuales de las reglas de desambiguación pueden servir de manera similar para la resolución de la subespecificación, con la diferencia que su parte ejecutiva debe cambiar un rasgo genérico por otro de explícito.

5.- Bibliografía

ARMSTRONG, S.; ROBERT, G. y P. BOUILLON: "Building a language model for POS tagging" (http://issco-www.unige.ch/tools).

BACH, C. ; SAURÍ, R.; VIVALDI, J. y M.T. CABRÉ (1997) "El Corpus de l'IULA: descripció", *Papers de l'IULA*, sèrie informes, 17, Barcelona: IULA, Universitat Pompeu Fabra.

BADIA, T.; PUJOL, M.; TUELLS, T.; VIVALDI, J.; DE YZAGUIRRE, Ll. y M.T. CABRÉ (en premsa) "IULA's LSP Multilingual Corpus: Compilation and Processing", Actes de la Conferència ELRA, Granada, maig de 1998.

DE YZAGUIRRE, L. , A. MATAMALA y T. CABRÉ (2000a) El lematizador "PALIC" del IULA (UPF), comunicación presentada al XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

DE YZAGUIRRE, L., S. TORNER Y A. MATAMALA (2000b) El tratamiento automático de las ambigüedades segmentales del castellano, comunicación presentada al XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

HINDLE, D. (1994) “A Parser for Text Corpora” en ATKINS, B. T. S. y A. ZAMPOLLI (eds.) “Computational Approaches to de Lexicon”, Oxford

MOREL, J.; TORNER, S; VIVALDI, J; DE YZAGUIRRE, Ll. y M.T. CABRÉ (1997) "El corpus de l'IULA: etiquetaris", *Papers de l'IULA*, sèrie informes, 18, Barcelona: IULA, Universitat Pompeu Fabra.

SILBERSZTEIN, M (1996) “Dictionnaires électroniques et analyse automatique de textes. Le système INTEX, Masson, Paris

TZOUKERMANN, É. y D. R. RADEV (1996) Using word class for part-of-speech disambiguation, en EJERHED, E. y I. DAGAN (eds.) “Fourth Workshop on very large Corpora, Copenhagen

VIVALDI, J.; DE YZAGUIRRE, Ll.; SOLÉ, X. y M.T. CABRÉ (1996) "Marcatge estructural i morfosintàctic del Corpus Tècnic amb l'estàndard SGML", *Papers de l'IULA*, sèrie informes, 1, Barcelona: IULA, Universitat Pompeu Fabra.