TURBIO
Sistema de extracción de información a partir de textos estructurados

Jordi Turmo Borràs
Dept. Lenguajes y Sistemas Informáticos
Universidad Politécnica de Cataluña
e-mail: turmo@goliat.upc.es
telf: 93-4017024

Palabras clave: Extracción de información

Resumen:

En este artículo se presenta un sistema para la extracción de información a partir de textos de dominio restringido. Nuestra propuesta, TURBIO, tiene dos funcionalidades: el aprendizaje de reglas de extracción a partir de una muestra significativa del corpus de origen y la ejecución de dichas reglas para la extracción de información del corpus. La principal aportación de TURBIO es, pues, que hace innecesaria la generación manual de reglas de extracción. Los resultados obtenidos en un experimento llevado a cabo para la extracción de información a partir de textos en el dominio micológico hacen pensar que la técnica de aprendizaje de reglas es transportable a otros dominios con un esfuerzo limitado.

1.- INTRODUCCIÓN Y MOTIVACIÓN

Habitualmente las bases de conocimiento se construyen manualmente mediante intervenciones de los expertos humanos en el dominio y la aplicación a desarrollar. Sin embargo, el coste que supone la construcción manual es enorme tanto en tiempo como en personal dedicado. Esto, unido a la existencia de fuentes textuales que contienen dicho conocimiento (bases de datos léxicas, diccionarios, enciclopedias y corpus, en orden decreciente en cuanto a la estructuración del conocimiento que aportan), hizo pensar en la posibilidad de automatizar la obtención de bases de conocimiento mediante la construcción de Sistemas Inteligentes Basados en Textos (SIBT).

Dentro de los SIBT se incluyen los Sistemas de Extracción Textual (SET), cuyo objetivo es obtener estructuras cognitivas que contengan el conocimiento extraído a partir de textos. Actualmente, los SET existentes tratan textos de dominio restringido, debido a la inexistencia de herramientas robustas de amplia cobertura que cubran los aspectos semánticos y pragmáticos de textos no restringidos. Ejemplos de SETs son SCISOR [Jacobs & Rau, 90], TACITUS [Hobbs et al, 91], SPARSER [McDonald, 92] y FASTUS [Hobbs et al, 93], en el dominio de las noticias periodísticas, NOMOS [Zarri,92] en el dominio legislativo, COBALT [Zarri,95] en el dominio financiero y nuestra propuesta, TURBIO.

TURBIO ha sido probado con textos de tipo enciclopédico sobre dominio micológico (fichas descriptivas cedidas por la Sociedad Micológica de Cataluña). Son tres las razones de dicha elección. Por un lado, la riqueza léxica, sobre todo adjetival, y la riqueza gramatical que ofrecen dichos textos. En los textos descriptivos sobre el dominio micológico la cobertura lingüística de los conceptos implicados es, en efecto, muy rica. Por otro lado, el dominio micológico es rico en conceptos difusos como en *sombrero de color algo amarillento* o en *de rojo claro a pardo rojizo*, en conceptos temporales como en *primero rojo pasando a pardo rojizo con la edad* y en conceptos multivalorados conjuntiva o disyuntivamente. Finalmente, resulta interesante la presencia de formas anafóricas, sobre todo elipsis como en *sombrero algo amarillento*, donde el rasgo **color** ha sido elidido. Dada la complejidad que supone el estudio de todos estos fenómenos ha sido necesario, para este trabajo, restringirnos a una cualidad, el color y restringir la morfología del hongo a las partes más relevantes que tuviesen esa cualidad. Como hemos visto en los ejemplos, el color es un rasgo altamente difuso, no sólo por su descripción de forma aislada, sino porque también puede aparecer como un intervalo, como una descripción temporal, como una conjunción o disyunción de valores y, además de aparecer generalmente como rasgo elidido, el conjunto de partes morfológicas aptas para tener color es elevado, por lo que el rasgo en cuestión es interesante para tratar las elisiones tanto del rasgo como de la parte morfológica que lo contiene. Es decir, pensamos que la restricción impuesta no atenta gravemente a la generalización de los resultados obtenidos. Estos resultados han sido utilizados para formar parte de la base de conocimiento del sistema experto KINOKO [Turmo,97] para la clasificación de setas.

Los siguientes apartados describen el sistema TURBIO. En el apartado 2 se explican las funcionalidades del sistema. El apartado 3 describe los módulos que componen tanto la arquitectura de TURBIO como el entorno en el que trabaja el mismo, centrándose en lo referente a la adquisición automática de reglas de extracción. Los resultados obtenidos en la ejecución de la extracción son presentados en el apartado 4. Finalmente, en el apartado 5 se describen las líneas de trabajo futuro.

2- FUNCIONALIDADES

Una de las mayores dificultades de los SETs reside en la definición de las reglas de extracción de información. Generalmente éstas vienen representadas por tuplas del tipo <palabra clave, conjunto de templetas>, donde la palabra clave denota un concepto del dominio y cada templeta asociada a ella contiene rasgos que modifican dicho concepto. El sistema, una vez analizado superficialmente el texto, busca iterativamente una palabra clave y sus modificadores y activa la templeta asociada a dicha palabra que maximiza la cobertura de los modificadores. Sin embargo, las reglas de extracción son definidas manualmente, con el consecuente coste temporal y humano que supone.

TURBIO propone la adquisición de reglas de extracción a través de un proceso de aprendizaje mediante un corpus de entrenamiento (conocimiento sobre el comportamiento del corpus). El conjunto de reglas obtenido será, posteriormente, utilizado para ejecutar la extracción de información.

Lo que se pretende exactamente es: a partir de una representación estructurada de un dominio (ej. el dominio micológico), extraer la información relevante contenida en textos en lenguaje natural que describen aspectos de dicho dominio. La unidad elemental a extraer corresponderá a tripletas del tipo <entidad atributo valor>.

Seguidamente se explica la metodología seguida por TURBIO para realizar ambas funcionalidades.

3.- ARQUITECTURA

El dominio (Fig.1) se representa utilizando un sistema de estructuras de rasgos tipificados, asociado a una taxonomía de tipos (LKB) [Copestake,92].

En lo que sigue, los ejemplos se refieren al dominio micológico y a la BD enciclopédica construida a partir de las fichas de descripciones micológicas (BDM).

Dicho entorno sirve a TURBIO tanto para la obtención de reglas de extracción como para la ejecución de la extracción que se realiza en el módulo NUCLEO mediante la aplicación de las reglas de extracción sobre todo el corpus.

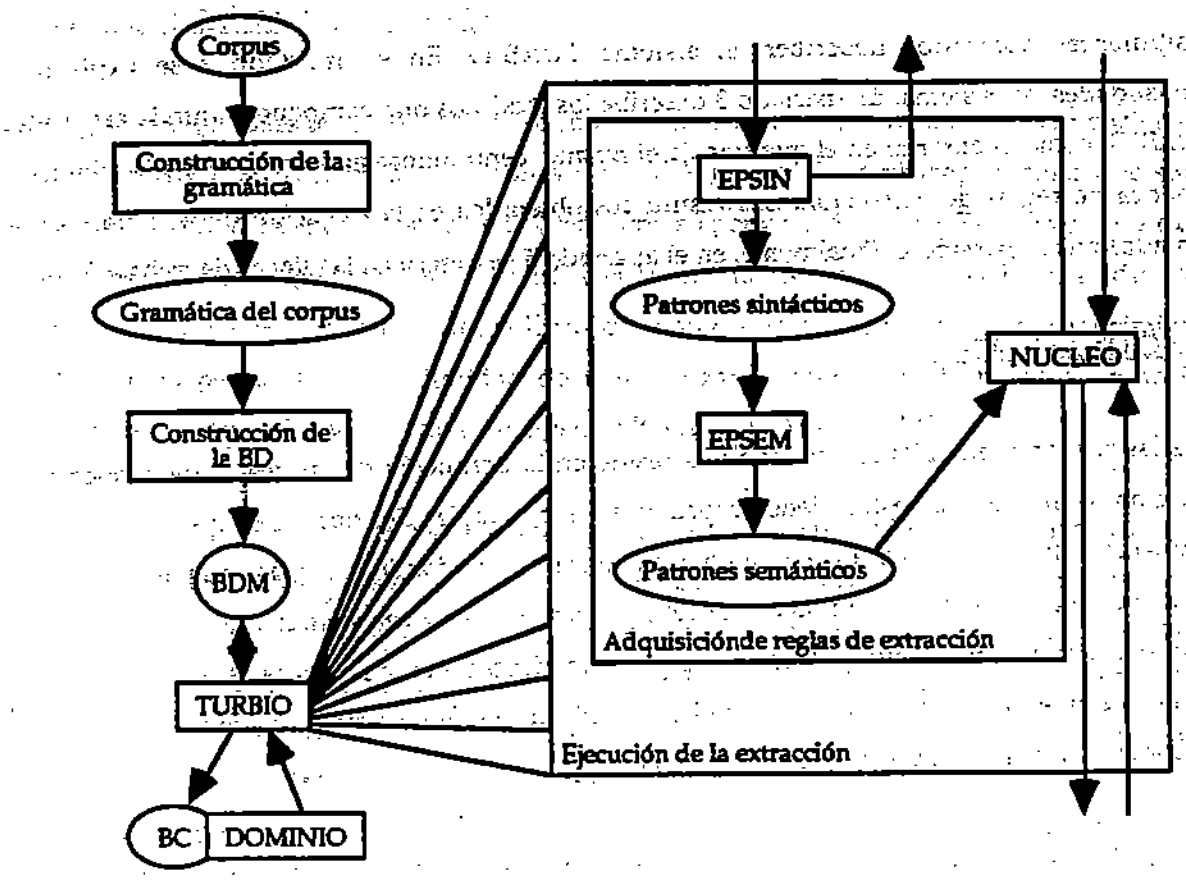

Fig.1.- Entorno y arquitectura de TURBIO

La BDM contiene información textual extraída del corpus. La construcción se realiza mediante un análisis estructural del texto con una gramática DCG de 81 reglas. Un ejemplo del resultado de tal análisis es el siguiente:

```
((Agaricus)
 (LNOM grec)
 (ENOM agarikón = de fong)
 (VARI campeter (L) Fr.)
 (LVAR llatí)
 (EVAR campester = campestre)
 ...
 (SENS 9)
 (TEXT castellà)
 (PRGF 1)
 (FRAS 1)
 (CONT Sombrero carnoso, hemisférico y luego aplanado, extendido y excepcionalmente
       hasta 20 cm, de color naranja vivo.)
 ...
 (SENS 20)
 (TEXT castellà)
 (PRGF 4)
 (FRAS 1)
 (CONT Comestible.)
 ...
```

La información contenida en los atributos CONT es la que sirve como corpus.

En cuanto al conocimiento del dominio[2], el sistema de tipos de la LKB proporciona la plataforma adecuada para representarlo. En la figura 2 se muestra una simplificación de los tipos de la holonimia del color y de las partes micológicas que tienen color, así como algunas restricciones de dichos tipos. El tipo *color-intervalo* tiene una restricción compuesta de dos rasgos, *de* y *a*, que permite almacenar información sobre el color expresado como intervalo - *de rojo a blanco amarillento* -. El tipo *fungi* representa a cualquier seta y por lo tanto se restringe a las partes más generales de ella (sombrero, carne y pie) que tienen color, es decir, que poseen una restricción sobre el pigmento. De esta manera, la formación de entradas de la base de conocimiento (BC) que debe extraer TURBIO se realizará mediante la unificación de las restricciones de tipos que han sido rellenadas con información en el proceso de extracción.

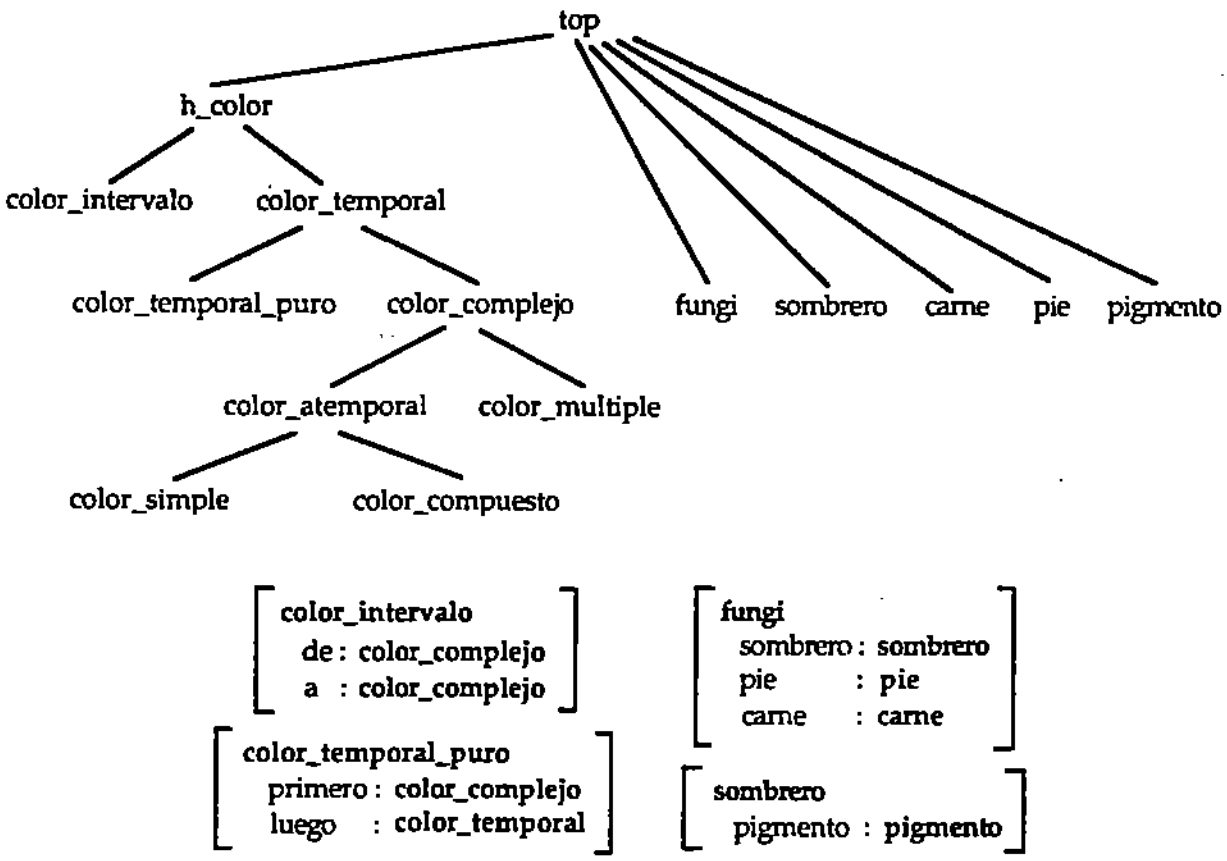

Fig.2.- Restricciones y sistema de tipos para el dominio micológico sobre el color

Respecto a la primera funcionalidad de TURBIO - adquisición de reglas de extracción - se produce en tres fases (Fig.1). En la primera de ellas (módulo EPSIN - Extractor de Patrones SINtácticos -) se analiza morfológicamente el corpus de entrenamiento para pasar después a analizarlo sintácticamente mediante el uso de una gramática superficial que produce segmentos parentizados de frases. Estos análisis serán incorporados a la BDM para su posterior tratamiento en la ejecución de la extracción. De cada uno de los segmentos se obtiene un patrón sintáctico sustituyendo cada palabra

[2] El dominio queda retringido al concepto *color* y las partes micológicas que lo pueden contener, tal y como se explicó en la introducción.

contenida en dicho segmento por una variable. Finalmente se eliminan, por una parte los patrones sintácticos menos frecuentes y aquellos que quedan totalmente cubiertos por otros mediante el uso de dos heurísticos y, por otra parte, aquellos patrones que no contienen etiquetas morfológicas relevantes (nombre, adjetivo y verbo). Ya en la segunda fase (módulo EPSEM - Extractor de Patrones SEMánticos) se generan los patrones semánticos para cada patrón sintáctico encontrado por EPSIN. Un patrón semántico es aquel que está dotado de significado dentro del dominio a tratar y es por ello que esta fase se inicia con la eliminación de aquellos patrones sintácticos cuyas variables no representan ninguna propiedad micológica, es decir, los conjuntos de palabras sustituidos por dichas variables no tienen significado dentro del dominio a tratar. Acto seguido se estudian estos conjuntos de palabras en cada patrón sintáctico no eliminado con el fin de generar patrones semánticos de manera que si un conjunto contiene palabras muy frecuentes en relación con las demás, cada una de ellas será candidata a formar parte de un patrón semántico, producto de sustituir la variable a la que representa la palabra por ella misma. Pero también es posible que formen un patrón semántico aquellas palabras, representadas por variables contiguas en un patrón sintáctico, que concatenadas tengan un significado dentro del dominio aunque sus frecuencias sean bajas[1] . Una vez obtenidos los posibles patrones semánticos se eliminan aquellos que no conciernen al dominio tratado y aquellos que simplemente son incoherentes gramaticalmente. En la tercera y última fase (módulo NUCLEO) se construyen las reglas de extracción.

A continuación se explica la metodología de cada uno de los módulos de TURBIO.

3.1.- Módulo EPSIN

El objetivo de este módulo es extraer los patrones sintácticos relevantes en el corpus, es decir, patrones que contemplan únicamente etiquetas sintácticas y son frecuentes en el corpus. Tres son las fases por las que pasa este proceso: el análisis y la desambiguación morfológica y el análisis superficial del corpus para la obtención de segmentos de frases sintácticamente analizados, que conforman un preproceso y, finalmente, la obtención de los patrones sintácticos relevantes a partir de dichos segmentos.

3.1.1.-Preproceso

Para el análisis morfológico se ha utilizado MACO [Acebo et al., 94] con un conjunto de 59 categorías simples. El resultado (que contiene un 46.7% de palabras ambiguas) fue posteriormente desambiguado utilizando RELAX [Padró, 96] El índice de corrección es bajo, 84.2%, aunque suficiente para nuestros propósitos. La causa hay que buscarla en las características del sublenguaje (alta proporción de adjetivos, uso de tecnicismos, nombres propios), mientras que el aprendizaje del etiquetador se llevó a

[1] Esto puede suceder puesto que las expresiones que encontramos en lenguaje natural dentro del corpus son muy variadas. Se puede dar el caso que una expresión no sea representativa en el corpus de entrenamiento para formar un patrón semántico, pero lo sea en la globalidad del corpus.

cabo con un corpus equilibrado.

Posteriormente se realiza el análisis superficial del corpus etiquetado. Para ello se ha utilizado una gramática superficial DCG compuesta por 109 reglas. De entrada, existe un proceso de partición de la frase en fragmentos capaces de ser analizados. El criterio de partición ha sido, simplemente, la utilización de signos de puntuación como separadores. Por ejemplo, la frase:

"Pie más corto, atenuado en la base, compacto, blanco y con frecuencia manchado de ocre."

genera la siguiente partición, previo análisis morfológico:

```
n("Pie") d("más") a("corto") zcoma(",")
a("atenuado") r0p ("en") j("la") n("base") zcoma(",")
a("compacto") zcoma(",")
a("blanco") c0c("y") u0v("manchado") r0p("de") a("ocre") zpunt(".")
```

Cada parte es analizada con la gramática superficial, que devuelve los posibles análisis de cada una de ellas. En el ejemplo, la primera partición daría como posibles análisis:

```
(n "Pie") (d "más") (a "corto") (zcoma ",")
(g_nom ((n "Pie"))) (d "más") (a "corto") (zcoma ",")
(g_nom ((n "Pie"))) (g_adj ((d "más") (a "corto"))) (zcoma ",")
(g_nom ((n "Pie") (g_adj ((d "más") (a "corto")))))) (zcoma ",")
```

En un postproceso se concatenan los análisis de las diferentes particiones, combinándolos ordenadamente entre sí y se da como único resultado la concatenación más larga de entre las más profundas. Así pues como resultado de nuestro ejemplo obtendríamos:

```
(g_nom ((n "Pie") (g_adj ((d "más") (a "corto")))))
(zcoma ",") (a "atenuado")
(sp ((r0p "en") (sn ((spec ((j "la"))) (g_nom ((n "base"))))))) (zcoma ",")
(a "compacto") (zcoma ",") (a "blanco") (c0c "y")
(sp ((r0p "con") (g_nom ((n "frecuencia"))))) (u0v "manchado")
(r0p "de") (a "ocre") (zpunt ".")
```

Posteriormente, el resultado es normalizado, mediante una sencilla gramática de transformaciones [Sager, 81] formada por solo 11 reglas, para tratar los siguientes casos: a) desglose de conjunciones nominales como por ejemplo *sombrero y pie amarillo claro* pasaría a ser *sombrero amarillo claro y pie amarillo claro* , b) el par adjetivo+nombre pasa a ser nombre+adjetivo como el caso de *rojo sombrero* por ser semánticamente equivalente a *sombrero rojo*, c) sustitución de abreviaturas del estilo *cm.* por su equivalente *centímetros* y d) Cambios de normalización tipográfica.

Una vez obtenidos los segmentos del análisis superficial EPSIN entra en la obtención de patrones sintácticos relevantes que se explica en el siguiente subapartado.

3.1.2.- Obtención de patrones sintácticos relevantes

Un patrón sintáctico puede interpretarse como un esquema resultado de abstraer un conjunto de segmentos o subsegmentos que responden al mismo árbol sintáctico. Por ejemplo los segmentos (g_nom ((n "sombrero") (a "blanco"))) y (g_nom ((n "forma") (a "ovoidal"))) generan el patrón sintáctico (g_nom ((n X) (a Y))), pero también los patrones (n X) y (a X) puesto que derivan de sus subsegmentos nominales y adjetivales.

Todo patrón sintáctico se asocia a dos informaciones diferentes: a) su conjunto ligado (CL) que informa sobre qué palabras, incluyendo su frecuencia en el corpus, son valores de cada variable del patrón y que para nuestro ejemplo sería (X ("sombrero" 1 "forma" 1) Y ("blanco" 1 "ovoidal" 1)) y b) su frecuencia absoluta que informa sobre el número de segmentos y subsegmentos que generan el patrón.

Como se puede observar en los patrones que derivan de los segmentos del ejemplo, los dos últimos están incluidos en el primero. Esto nos permite definir una relación entre patrones, a la que llamaremos *relación de cobertura* , de la siguiente manera: sea P el conjunto de patrones sintácticos,

$$r,t \in P \text{ , t cubre a r } \Leftrightarrow \text{ r cubierto por t } \Leftrightarrow t \sqsupseteq r \Leftrightarrow \text{ t incluye a r}$$

Para poder clarificar cual es el objetivo de esta fase necesitamos establecer antes una clasificación de los patrones sintácticos desde el punto de vista de su derivación a partir de los segmentos analizados.

Podemos, pues, clasificar los patrones sintácticos en aquellos derivados de segmentos y aquellos derivados de subsegmentos (Fig.3). La intersección de ambos conjuntos no es nula puesto que, por ejemplo, el patrón (a X) deriva del segmento que se obtienen al analizar la porción de frase *rojo* y del subsegmento resultado del análisis de *rojo* en la porción de frase *su color rojo*.

La razón de nuestro objetivo es que de todos los patrones sintácticos se pueden despreciar aquellos que son de baja frecuencia absoluta y, en principio, aquellos que derivan de subsegmentos, puesto que la información que podríamos obtener con ellos es menos específica que la que contienen los segmentos. Sin embargo existen segmentos que contienen información relevante sobre el dominio pero no son frecuentes en el corpus, por lo tanto sus patrones serán despreciados. Esto haría perder dicha información relevante. Por consiguiente, se deben tener en cuenta tanto lo patrones derivados de segmentos de alta frecuencia como los patrones derivados de subsegmentos con información relevante contenidos en segmentos que generan patrones de baja frecuencia.

Para obtener los patrones relevantes es necesario estudiar sólo los pares de patrones de alta frecuencia absoluta relacionados bajo cobertura, de manera que serán relevantes aquellos que no estén cubiertos por otros en todas sus apariciones en el corpus.

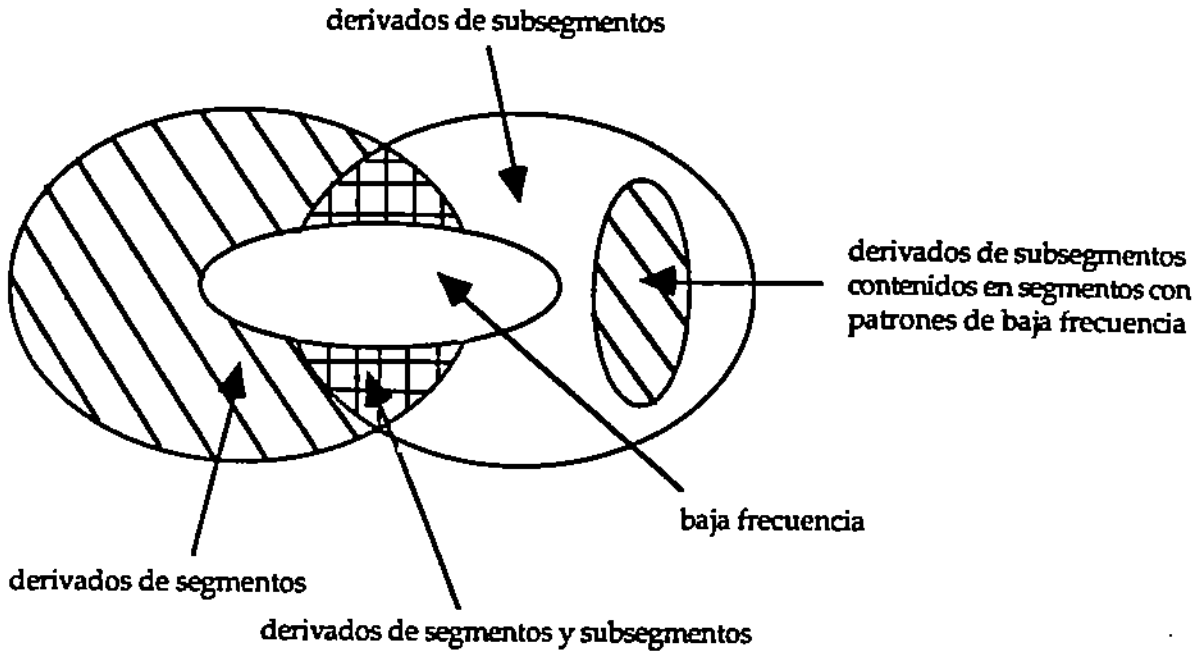

Fig.3.- Clasificación de patrones sintácticos según su derivación

Se han codificado los patrones de forma sencilla asociando un número primo a cada etiqueta categorial y asociando a cada patrón el producto de los números primos que representan a cada una de las categorías contenidas en el patrón (codificación ∂). Por ejemplo, el patrón r = (g_nom ((n X)(a Y))) tiene como codificación $\partial(r) = 30$ si se tiene las asociaciones g_nom=2, n=3 y a=5. Dicha codificación permite definir la relación de equivalencia $=_\partial$ entre patrones como sigue:

$$r, s \in P,\ \partial(r) = \partial(s) \Rightarrow r =_\partial s$$

Gracias a esta nueva relación podemos formar el conjunto cociente $P / =_\partial$ que proporciona conjuntos de patrones ∂-equivalentes (∂-conjuntos). Entre los elementos del conjunto cociente se pueden establecer una relación de ligadura si algún patrón del primer elemento puede cubrir a otro del segundo elemento (∂-conjuntos ligados). La ligadura se verifica de la siguiente manera:

$$\partial_a, \partial_b\ \partial\text{-conjuntos ligados} \Leftrightarrow \partial_a \bmod \partial_b = 0$$

En la figura 4 podemos ver un ejemplo en el que los conjuntos 30 y 210 quedan ligados debido a que t cubre a r.

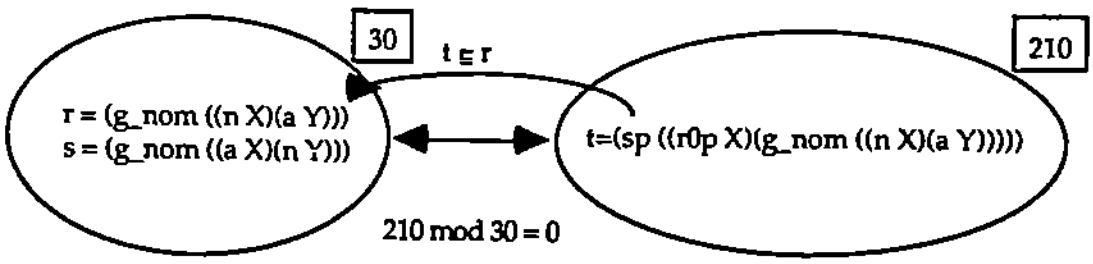

Fig.4.- ∂-conjuntos ligados

Así pues el método para obtener lo patrones sintácticos relevantes pasa por las siguientes etapas:

a) Obtención de los ∂-conjuntos
b) Eliminación de patrones de baja frecuencia absoluta ($W_i < 10$)
c) Reducción de los ∂-conjuntos. Si definimos como frecuencia residual de un patrón como la diferencia entre su frecuencia absoluta y las frecuencias absolutas de todos aquellos que lo cubren, eliminaremos aquellos patrones que tengan una frecuencia residual nula, es decir, que ya estén cubiertos por otros patrones más precisos.
d) Eliminación de patrones que no contienen etiqueta categorial nominal, verbal o adjetival.

Una vez acabado el proceso los patrones sintácticos relevantes son utilizados por el módulo EPSEM.

3.2.- Módulo EPSEM

El objetivo de este módulo es extraer patrones semánticos a partir de los sintácticos relevantes. Un patrón semántico se define como un patrón sintáctico dotado de significado dentro del dominio a tratar. Como dijimos en el apartado 2 las unidades a extraer son de la forma < parte-miceto, atributo-de-parte, valores>. Podemos clasificar los patrones semánticos en genéricos y específicos, dependiendo de la información que contienen. Los primeros son aquellos que no contienen información sobre el atributo de la parte, a diferencia de los segundos, que sí que la contienen ya sea explícita o implícitamente. Otra manera de clasificar los patrones semánticos es según el modo en que han sido generados. Encontramos, entonces, patrones semánticos simples y compuestos.

La extracción de patrones semánticos simples se realiza a partir del análisis de los CLs asociados a los patrones sintácticos. En primer lugar se eliminan aquellos patrones sintácticos relevantes de CLs inválidos para el dominio, es decir, cuyos CLs no revelan que contengan información válida para el dominio a tratar. Por ejemplo, el patrón sintáctico:

(sn ((spec ((j X))) (g_nom ((n Y)))))

típico de expresiones como *el sombrero*, pertenece a esta clase de patrones, por lo que puede ser eliminado. En segundo lugar se generan patrones semánticos combinando palabras de alta frecuencia relacionadas con cada variable de un patrón sintáctico no eliminado. Para el patrón:

(sp ((r0p X) (g_nom ((n Y) (n Z)))))

se generan los patrones semánticos:

p1=(sp ((r0p "a") (g_nom ((n X) (n Y)))))
p2=(sp ((r0p "a") (g_nom ((n "veces") (n X)))))

siendo p1 típico de expresiones como *a rojo amarillento*. Sin embargo, de todos estos patrones

semánticos generados existen algunos que no tiene frases representantes que hagan referencia a conceptos del dominio y que, por lo tanto, deberán ser eliminados. Es el caso del patrón p2. Analizando su CL se observa que no existe ninguna palabra asociada a la variable X que haga posible que p2 sea un patrón semántico válido.

Por último, se generan patrones semánticos por esquemas significantes. Pueden existir combinaciones de palabras relacionadas con cada variable de un patrón sintáctico que tengan un significado importante dentro del dominio y no hayan generado un patrón semántico debido a que dichas palabras son de baja frecuencia. Es el caso de:

(sp ((r0p X) (g_nom ((n Y) (n Z)))))

que, de esta manera, genera

p3=(sp ((r0p "de") (g_nom ((n "color") (n X))))).

Una vez extraídos los patrones semánticos simples, EPSEM pasa a obtener los compuestos a partir de los primeros, ya sea combinándolos entre ellos, como el caso de p3 + p1 para generar un patrón capaz de soportar expresiones del color como intervalo y combinando patrones fijos con simples como el caso (verbal ((x "se) (v0v "vuelve"))) + (a X) típico de expresiones que denotan viraje del color como *se vuelve negruzco con la edad* . Finalmente se eliminan aquellos patrones simples que no tienen significado sin combinar, como p1.

Las dos clasificaciones pueden ser relacionadas. Mientras que todos los patrones semánticos compuestos son patrones semánticos específicos (ej.: (sp ((r0p "de") (g_nom ((n "color") (n X))))) + (sp ((r0p "a") (g_nom ((n X) (n Y)))))) informa sobre el rasgo color expresado como intervalo) existen patrones semánticos simples que son específicos (ej.: (sp ((r0p "de") (g_nom ((n "color") (n X))))) que informa del rasgo color) y otros que son genéricos (ej.: (a X) que informa sobre el valor de algún rasgo).

3.3.- Módulo NUCLEO

En este último módulo de TURBIO se obtienen las reglas de extracción mediante dos submódulos, CPS y GRE (Fig.5).

El CPS clasifica los patrones semánticos en una jerarquía de prioridades siguiendo dos criterios: 1) los patrones semánticos específicos son más prioritarios que los genéricos y 2) sean p y q patrones semánticos, si p⊏q entonces p es más prioritario que q, es decir, a mayor longitud del patrón, mayor prioridad.

Finalmente el GRE genera las reglas de extracción, de manera que la condición de cada una de ellas es uno de los patrones semánticos y la acción es el método de extracción que le corresponde a dicho

patrón. Actualmente existen cinco métodos de extracción, uno para cada tipo de patrón genérico (los que informan sobre la parte del miceto y sobre el valor del rasgo elidido y los que únicamente informan sobre el valor elidiendo tanto la parte como el rasgo) y uno para cada tipo de patrón específico (los que informan de la parte, el rasgo y el valor, los que informan del rasgo y el valor y los que informan del valor, quedando implícito el rasgo).

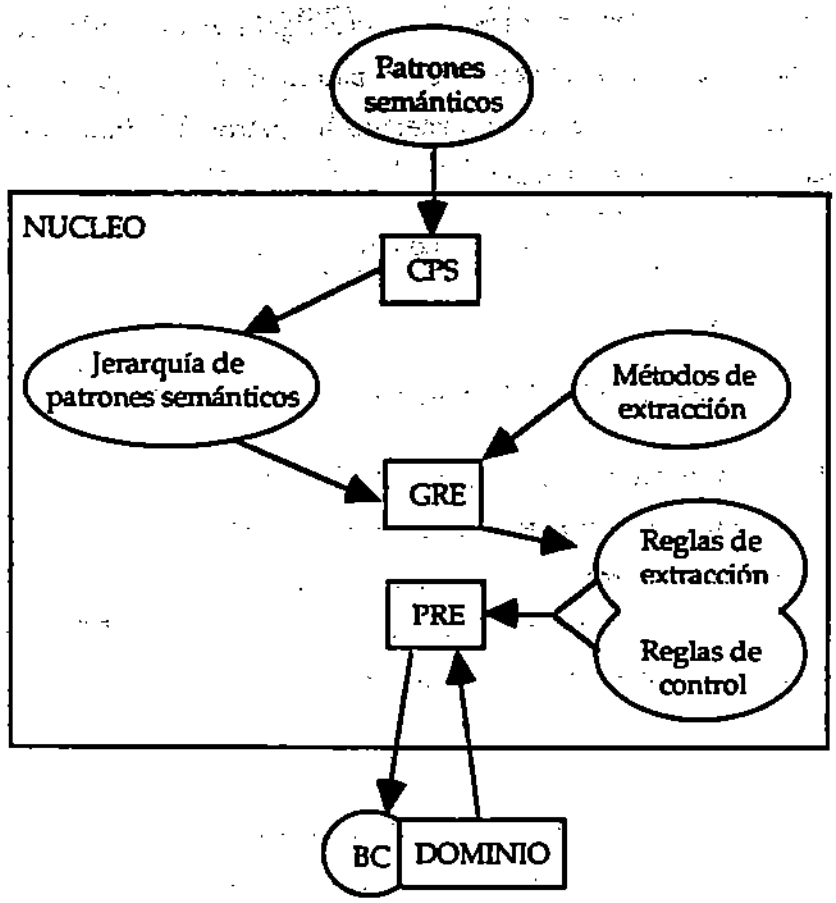

Fig.5.- Arquitectura del módulo NUCLEO de TURBIO

Posteriormente se ejecuta la extracción mediante la aplicación de dichas reglas sobre el corpus utilizando PRE [Ageno et al., 94] .

Cada una de las reglas se clasifica en prioridades para su ejecución en PRE adoptando la misma prioridad que tiene su patrón semántico en la jerarquía obtenida por el CPS. Así pues, un ejemplo de regla es:

```
(rule general23
  ruleset general
  priority 2
  control one
  (actual ^id ?id ^a (*r1 (a ?w1) *r2))
  (actual ^id ?id ^a ?arbre)
  (aux ^fss ?fss-actual)
  (context ^id ?id-1 ^actual ?a ^frases-ok ?f ^cua *?)
  ->
  (?tipus-ok := (contextualitzar (gethash (intern ?w1) la) ?a ?f))
```

```
(?fs-ok :=
  (if ?tipus-ok
     (unificar-tipus-valor (copy-type-entry (cadar ?tipus-ok))
       (copy-type-entry (car (gethash (intern ?w1) la)))
      ?w1 ?arbre)))
(?* := (when ?fs-ok
          (put-aux ?fs-ok ?fss-actual)
            (modify-wm 'actual nil  (list 'actual (list '^id ?id '^a (list *r1 *r2)))))))
```

regla de prioridad 2 por la cual si un segmento del corpus analizado unifica el patrón semántico (a X) se contextualiza el valor de X para encontrar la parte y el rasgo correcto y se obtiene la estructura de rasgos correspondiente.

4.- APLICACIÓN Y RESULTADOS

TURBIO ha sido probado utilizando 150 fichas descriptivas de setas cedidas por la Sociedad Catalana de Micología con un total de 21609 palabras correspondientes a 2991 formas.

Con el módulo EPSIN se obtuvieron 159 patrones sintácticos en la obtención de los ∂-conjuntos incluyentes. De ellos se eliminaron 73 que tenían frecuencia absoluta menor que 10, quedando un total de 86, de los cuales 2 fueron despreciados en la etapa de reducción de los ∂-conjuntos. Finalmente, de los 84 restantes 35 fueron eliminados por ser patrones superfluos en etiqueta categorial. En total se obtuvieron 49 patrones sintácticos relevantes.

Ya en el módulo EPSEM, se encontraron 4 patrones sintácticos relevantes con CL's inválidos. De los 45 restantes sólo 28 hacían referencia al color y de ellos se extrajeron 53 patrones semánticos simples y 68 compuestos, eliminando, posteriormente, 7 simples por no tener significado sin combinación.

Así pues, en total se encontraron 114 patrones semánticos válidos. Todos ellos serían utilizados precisamente para la obtención de reglas de extracción por el módulo NUCLEO de TURBIO, tal y como se explica en el siguiente apartado.

El resultado de la extracción fue de un 48.30% de cobertura y un 87.14% de precisión, distribuyéndose el 51.70% de casos no cubiertos de la siguiente manera: un 27.75% de casos fueron debidos a la inexistencia de regla (16.26% parcialmente extraídos), un 12.46% fueron debidos a errores de etiquetaje morfológico y el 11.49% restante de casos fueron debidos a la no obtención del coreferente, o bien porque el tipo estaba elidido y no se encontró en el contexto (6.22%), o bien porque el valor estaba referenciado a otra parte del miceto mediante una comparación (5.27%).

Los resultados son difíciles de evaluar ya que no existe (como en el caso de MUC para el inglés) un patrón válido de comparación. Si los comparamos con los resultados de MUC los podemos considerar

buenos teniendo en cuenta que sólo extraemos un atributo, el color, pero que la dificultad es bastante superior a la que presentan los atributos a extraer en los modelos MUC. En cualquier caso son esperanzadores de cara al futuro.

5.- CONCLUSIÓN Y TRABAJO FUTURO

En este trabajo se ha presentado un sistema para la extracción de conocimiento a partir de textos de dominio restringido utilizando reglas de extracción aprendidas a partir de una muestra significativa del corpus de origen.

El sistema ha sido aplicado al dominio micológico obteniendo unos niveles de precisión y cobertura aceptables.

Las líneas de trabajo futuro se canalizan a la mejora del resultado de la extracción obtenido por TURBIO. Como se ha visto, TURBIO obtuvo un 87.14% de rasgos correctamente extraídos. Sin embargo, era posible aumentar el porcentaje utilizando herramientas más potentes para obtener un análisis morfológico más correcto y obtener los coreferentes de expresiones cuyo patrón sintáctico es del tipo <lista-valor> o <rasgo>+<lista-valor>. Está previsto mejorar la cobertura del analizador morfológico utilizado (MACO), añadiendole más léxico general y específico del dominio a tratar y probar otras técnicas de desambiguación morfológica que permitan restringir, mediante reglas, ciertas etiquetas para palabras que cumplan algunas características (por ejemplo en el corpus utilizado no existe ningún verbo que esté en primera persona, sin embargo, algunas palabras son etiquetadas como verbo por el hecho de contener una terminación igual a la primera persona de una de las posibles conjugaciones verbales). Se puede, asimismo, utilizar información semántica léxica disponible para mejorar la eficacia de los desambiguadores [Padró, 96] de forma que se adapten al tratamiento de corpus (y sublenguajes) específicos. Por otra parte se pretende utilizar un analizador tipo CHART bidireccional dirigido por islas para mejorar la obtención de referentes, de forma que el análisis inicial pueda, incrementalmente, ampliarse a medida que las necesidades de referencia lo precisen.

Interesa, por otra parte, obtener resultados utilizando otros tipos de dominios que pudieran ser más sencillos respecto a sus expresiones lingüísticas (como el dominio legislativo). De esta manera se puede plantear un estudio comparativo de la respuesta de TURBIO a diferentes tipos de dominio restringido. Los sistemas de extracción actuales han sido probados con dominios legislativos y periodísticos, por lo que sería posible hacer un estudio comparativo entre esos sistemas y TURBIO.

La posibilidad de tratar corpus de diferentes lenguas (el inglés, utilizando el patrón MUC para evaluar resultados) o textos bilingües no alineados pretende, también, aumentar el porcentaje de aciertos del método. Existen fenómenos lingüísticos que son más comunes en una lengua que en otra, fenómenos que no permitirían al sistema la extracción de algún rasgo relevante en una lengua concreta.

Esta posibilidad permite que si alguna información no ha sido encontrada en un texto escrito en una lengua, pueda ser obtenida a partir del análisis de otro texto escrito en otra lengua.

En este sentido se han realizado pruebas iniciales con textos bilingües débilmente alineados y con la utilización de BDs multilingües (EuroWordNet).

Bibliografia

[Acebo et al., 94] S. Acebo, A. Ageno, S. Climent, J. Farreres, L. Padró, F. Ribas, H. Rodríguez, O. Soler. "MACO: Morphological Analyzer Corpus-Oriented". Report LSI-94-30-R UPC

[Ageno et al., 94] A. Ageno, I. Castellón, M.A. Martí, G. Rigau, H. Rodríguez, M. Taulé, F. Verdejo. "TGE: Tlink Generation Environment". Report LSI-94-9-T UPC

[Copestake, 92] A. Copestake. "The Acquilex LKB: Representation issues in semi-automatic acquisition of large lexicons". En Proceedings of 3rd Conference on Applied Natural Language Processing, Trento, Italy. Esprit BRA-30-30 Acquilex wp n.036.

[Hobbs et al, 91] J.R. Hobbs, D.E.Appelt, J. Bear, M. Tyson, D. Magerman. "Robust Processing of Real-World Natural-Language Texts". Artificial Intelligence Center. SRI International. Menlo Park, California.

[Hobbs et al., 93] J.R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Tyson. "FASTUS: A System for Extracting Information from Text" ARPA-93

[Jacobs, 92] P.S. Jacobs. Editor."Text-Based Intelligent systems". Lauwrence Erlbaum Associates, Publishers. Hillsdale, New Jersey.

[Jacobs & Rau, 90] P.S. Jacobs, L.F. Rau. "SCISOR: Extracting Information from Online News". Communications of the ACM, 33 (11), 88-97.

[McDonald, 92] D.D. McDonald. "Robust Partial-Parsing Through Incremental Multi-Algorithm Processing". Brandeis University and Content Technologies, Inc. 14 Brantwood Road, Arlington, MA 02174.

[Padró, 96] L. Padró. "POS Tagging Using Relaxation Labelling"
COLING'96. Copenhagen, Denmark

[Sager, 81] N. Sager. "Natural Language Information Processing". Addison-Wesley Publishing Company. Advanced Book Program Reading, Massachusetts. 1981.

[Turmo,97] J. Turmo. "KINOKO: Un sistema experto para la clasificación de micetos basado en hechos". Tesis de licenciatura en informática. UPC.

[Zarri,92] G.P. Zarri. "Encoding the Temporal Characteristics of the Natural Language Descriptions of (legal) Situations", En "Expert Systems in Law" de A. Martino. Ed. Amsterdam: Elsevier Science Publisher.

[Zarri,95] G.P. Zarri. "Knowledge Acquisition from Natural Language Documents for Large Knowledge Bases". Centre National de la Recherche Scientifique EHESSCAMS. Paris.