Del Texto a la Información (versión ampliada)

J. Atserias, N. Castell, N, Català, H. Rodríguez y J. Turmo

Dept. Llenguatges i Sistemes Informàtics

Universitat Politècnica de Catalunya

Campus Nord - Mòdul C6

c/ Jordi Girona Salgado 1-3, 08034 Barcelona

{batalla, castell, ncatala, horacio, turmo}@lsi.upc.es

Resumen

Las aplicaciones informáticas centradas en el Tratamiento de la Lengua (TL) han experimentado en los últimos años un notable auge sobre todo en el ámbito del acceso a la información textual no restringida (ni codificada). En este contexto están adquiriendo importancia creciente los sistemas de extracción de información a partir de textos no restringidos. Nuestra propuesta presenta dos vertientes: por una parte presentamos un entorno de extracción de información, es decir, una aplicación concreta de TL. Por otra parte describimos como en este entorno se integran diferentes módulos que abordan diferentes problemas de TL en castellano (análisis léxico y morfológico, análisis sintáctico superficial, aprendizaje de reglas de extracción, extracción de información, etc.) que potencialmente podrían utilizarse en otras aplicaciones.

Abstract

In the last years, software applications centered on Natural Language Processing (NLP) have experimented a great significance, especially access to neither unrestricted, nor codified, textual information. In this context, information extraction systems from unrestricted texts, have been acquired a growing interest. Our proposal is twofold: on one side we present an information extraction environment, i.e. a specific NLP application. On the other side we describe how different modules, dealing different issues on Spanish NLP (lexical and morphological analysis, shallow syntactic analysis, learning text extraction rules, information extraction, etc.), are integrated in it, and how these modules could be used in new applications.

Del Texto a la Información

J. Atserias, N. Castell, N, Català, H. Rodríguez y J. Turmo

Dept. Lenguajes y Sistemas Informáticos

Universidad Politécnica de Cataluña

Campus Nord - Módulo C6

c/ Jordi Girona Salgado, 1-3

08034 Barcelona

{batalla, castell, ncatala, horacio, turmo}@lsi.upc.es

Resumen

Las aplicaciones informáticas centradas en el Tratamiento de la Lengua (TL) han experimentado en los últimos años un notable auge sobre todo en el ámbito del acceso a la información textual no restringida (ni codificada). Aplicaciones como los sistemas de recuperación, formateo, filtrado o encaminado de información (por ejemplo en el área de internet) están en el ánimo de todos. En este contexto están adquiriendo importancia creciente los sistemas de extracción de información a partir de textos no restringidos.
Este auge ha dado lugar a la aparición de una nueva disciplina, la Ingeniería Lingüística, que aborda todos los aspectos (técnicas, métodos, herramientas, recursos) que conducen a la construcción de aplicaciones basadas en el TL.
Nuestra propuesta se enmarca en esta doble corriente: por una parte presentamos un entorno de extracción de información, es decir, una aplicación concreta de TL. Por otra parte describimos como en este entorno se integran diferentes módulos que abordan diferentes problemas de TL en castellano (análisis léxico y morfológico, análisis sintáctico superficial, aprendizaje de reglas de extracción, extracción de información, etc...) que potencialmente podrían utilizarse en otras aplicaciones.

1. Introducción

Las aplicaciones informáticas relacionadas con en el Tratamiento de la Lengua (TL) han experimentado en los últimos años un notable auge centrado básicamente en el tratamiento de información textual. Temas como la recuperación de información multilingüe, el filtrado y encaminado de información en función del perfil e intereses del usuario, los entornos de ofimática textual, la extracción automática de información a partir de textos relevantes o la traducción automática de textos no restringidos configuran un mercado amplísimo para el desarrollo de aplicaciones de TL. Por otra parte, el estado actual de la tecnología, tanto en el aspecto del hardware, como del software -sistemas de gestión de información léxica, analizadores, sistemas de interpretación semántica- como del lingware[1] -diccionarios y lexicones generales y terminológicos, gramáticas de amplia cobertura, ontologías lingüísticas y conceptuales- permite el desarrollo de muchas de las aplicaciones anteriormente mencionadas. El común denominador de estas aplicaciones es el tratamiento masivo de textos no restringidos, escritos en lengua natural y producidos para la utilización humana. Las dos aplicaciones emblemáticas en esta línea son la recuperación de información y la extracción de información. Aunque nuestro trabajo se centra en la segunda de ellas es importante precisar sus diferencias ya que a menudo se confunden.

Los sistemas de recuperación de la información ("information retrieval"), SRI, tienen como misión la selección de documentos (o textos) potencialmente relevantes para las necesidades del usuario de entre un conjunto normalmente muy alto de documentos posibles. Normalmente las necesidades del usuario se establecen en forma de una consulta ("query"). Los diferentes SRIs permiten formas más o menos complejas de expresión de las consultas pero raramente van más allá de la utilización de una lista de palabras, posiblemente relacionadas mediante conectivos lógicos. Las técnicas utilizadas en los SRI suelen tener una base estadística aunque el papel del TL es creciente (lematización de los términos del índice y/o de la consulta, generalización mediante el uso de ontologías, traducción a otras lenguas, etc...) .

Los sistemas de extracción de información (SEI) y otras aplicaciones de la misma familia como los formateadores de información o los sistemas de producción de resúmenes tienen unos objetivos diferentes a los SRI aunque buena parte de las técnicas básicas de TL que utilizan unos y otros coinciden. En un SEI lo que se pretende es extraer la información relevante contenida en un texto para incorporarla a una estructura de información (a menudo una base de conocimiento o una base de datos) que permita su proceso posterior. En un SEI es imprescindible la prescripción del esqueleto, formato o esquema sobre el que proyectar la información presente en el texto. Este esquema se

---

[1]Este último neologismo, creado por analogía a los anteriores, intenta recoger los aspectos relativos a los recursos lingüísticos que son utilizados por los sistemas de TL.

utiliza como fuente de información fundamental durante el proceso de extracción. Así, en un ejemplo bien conocido, las conferencias MUC (Message Understanding Conference), el esquema básico de la base de conocimiento está constituído por un sistema de objetos o clases ("frames") para cada uno de los cuales se han definido una serie de atributos susceptibles de recibir información de un tipo determinado. Las diferentes clases pueden estar relacionadas por diferentes tipos de relaciones. Tanto el conjunto de clases como el de atributos y relaciones suele estar prescrito y constituir un conjunto cerrado. La acción del SEI consiste entonces en la localización en el texto de posibles ejemplares de algunas de las clases prescritas, en el relleno de los atributos que aparezcan, explícitamente o implícitamente, citados en el texto y en el establecimiento de posibles relaciones entre dichos ejemplares. Es posible asimismo la extracción incremental de información a partir de varios documentos. A diferencia de lo que se pretendía en los SRIs, indiciar los documentos para permitir las consultas posteriores, lo que aquí se pretende es extraer directamente la información existente en los textos. La consulta posterior no se llevará a cabo contra éstos sino contra la estructura de información resultante.

2. Los sistemas de extracción de información

Evidentemente la estructura de los SEI es muy variada aunque los componentes básicos sean los mismos. Una arquitectura típica podría constar de los siguientes módulos:

- Segmentación del texto. Obtención de los fragmentos relevantes. Posiblemente detección de la lengua en que está escrito.
- Análisis léxico. Detección de las unidades léxicas (palabras, términos, locuciones, etc...). Consulta a diccionarios (generales o específicos del dominio), lexicones y bases de datos terminológicas. Detección de nombres propios, lexías, fechas, números, fórmulas, siglas, etc…
- Desambiguación de la categoría gramatical ("POS tagging").
- Lematización de las unidades léxicas.
- Desambiguación semántica ("word sense disambiguation"), es decir, asignación a cada unidad de su acepción correcta en su contexto de aparición.

- Análisis sintáctico, normalmente de tipo superficial y/o parcial.
- Análisis referencial. Búsqueda de los referentes conceptuales de las unidades léxicas. Identificación de las co-referencias (p. ej. referentes de los pronombres o de descripciones abreviadas de un objeto).
- Aplicación de las reglas de extracción de información.
- Incorporación al Sistema de Representación de la Información.

La visión actual es que estas diferentes etapas pueden ser desarrolladas y evaluadas independientemente, a diferencia de la aproximación tradicional que tendía a sugerir que la interdependencia entre todos los módulos no permitía su separación y uso en cascada (pipe-line). En este marco cabe destacar la aparición de arquitecturas que permiten la reutilización de estos módulos para las diferentes tareas que se engloban dentro del TL (dando lugar a la aparición de la Ingeniería de la lengua).

Uno de los más completos entornos de integración es GATE[2] (General Architecture for Text Engineering), que ha sido utilizado en este trabajo dentro del proyecto ITEM[3]. GATE es un entorno gráfico desarrollado en la Universidad de Sheffield, basado en la Arquitectura Tinman, que ha sido definida por el grupo de trabajo Tipster en el marco del DARPA.

3. Integración de herramientas para el TL en castellano

La función de la arquitectura GATE es la integración de técnicas de procesamiento lingüístico que pueden llevarse a cabo en ella. La integración permite la construcción de sistemas destinados a una tarea determinada, a partir de una serie de módulos que realizan un proceso lingüístico especifico. Estas tareas pueden pertenecer a niveles diversos del TL, desde anotadores y analizadores sintácticos hasta sistemas de extracción de información o de traducción automática.

---

[2] http:/ / www.dcs.shef.ac.uk/ research/ groups/ nlp/ gate
[3] (TIC96-1243-C03).

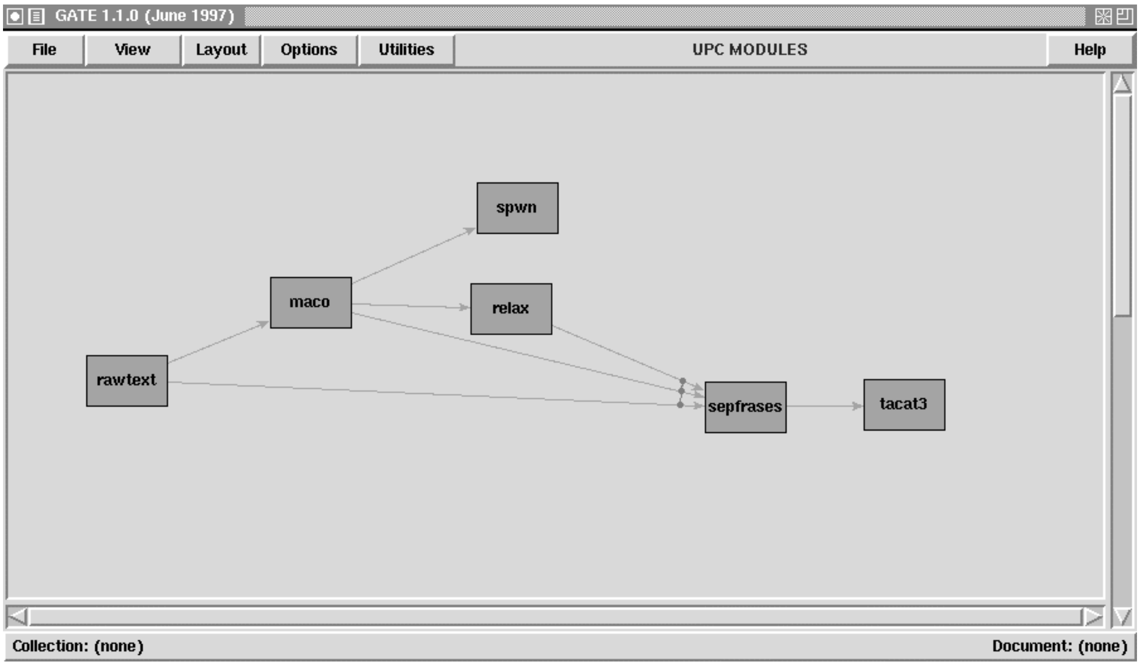

Fig. 1.- Módulos GATE para el análisis sintáctico del castellano

Nuestro uso de la arquitectura ha sido añadir nuevos módulos (Fig. 1) desarrollados en nuestro entorno de trabajo a los proporcionados por el sistema. A continuación los describimos.

3.1 TACAT (TAgged Corpus Analyser Tool)

TACAT [1] es un analizador sintáctico basado en charts[4] construido para proporcionar robustez y flexibilidad.

La entrada al analizador es un texto previamente etiquetado y desambiguado total o parcialmente (cada palabra va seguida de una o varias etiquetas). Para el etiquetado disponemos de un analizador morfológico (MACO [2]) y un desambiguador morfológico basado en relajación por restricciones (RELAX [3]). El conjunto de etiquetas puede ser definido libremente por el usuario[5], teniendo en cuenta que las etiquetas del texto y de la gramática han de ser compatibles. TACAT puede aceptar como entrada no sólo corpus etiquetado sino texto semi-analizado, de manera que se puede ir analizando el corpus en etapas sucesivas y revisando a cada paso el resultado.

---

[4] El analizador sigue una estrategia ascendente. Incorpora un mecanismo de tratamiento de las producciones épsilon y utiliza heurísticos para seleccionar el mejor árbol de análisis.

[5] El etiquetado que actualmente utilizamos sigue la codificación de Parole.

TACAT permite la actuación en cascada de varias gramáticas de contexto libre. Un conjunto de directivas permite controlar la actuación del analizador y la producción del árbol de análisis (categorías a omitir, aplanado del árbol , etc...).

El análisis que TACAT lleva a cabo es parcial, de forma que el resultado del análisis puede ser una serie de árboles sin ligar. TACAT dispone, además, de un entorno gráfico de manipulación de los árboles de análisis guiado por la gramática.

En el marco de ITEM se ha integrado, asímismo, en GATE un analizador de charts probabilísticos guiado por islas.

3.2 TURBIO

TURBIO [4] es un sistema de extracción de información a partir de textos de dominio restringido[6].

Para poder extraer información de textos es necesario que el sistema sepa qué conocimiento del dominio es relevante (ontología del dominio) y como extraer la información relevante de los textos (reglas de extracción). Típicamente ambos conocimientos son generados manualmente. TURBIO permite, sin embargo, liberar al experto de diseñar el conjunto de reglas de extracción adquiriéndolas automáticamente a partir de un corpus de aprendizaje, de forma que sólo se requiera intervención humana en el proceso de creación de la ontología.

---

[6] Los ejemplos que aquí presentamos pertenecen al dominio de la micología.

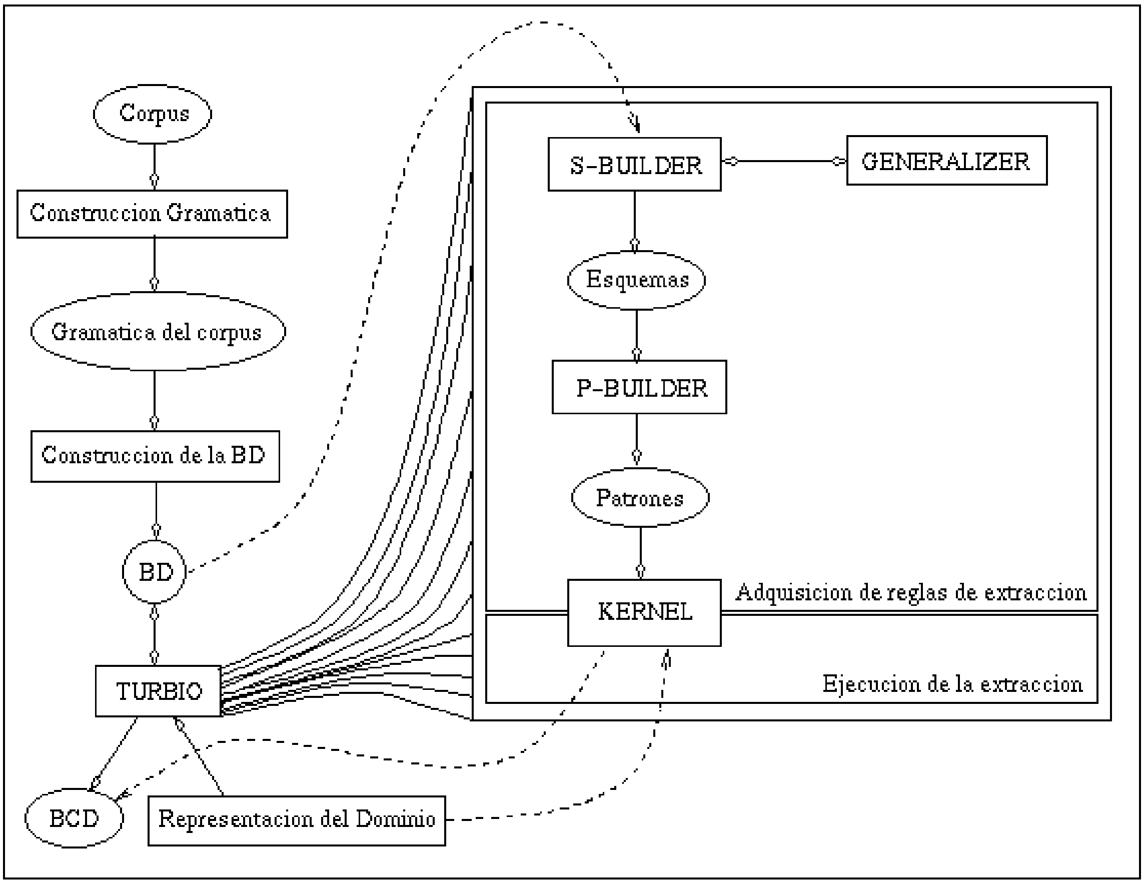

Fig. 2.- Entorno y arquitectura de TURBIO

En la figura 2 se muestra la arquitectura de TURBIO y su entorno. Dicho entorno utiliza una gramática que describe la estructura de los documentos para construir una base de datos (BD) que contiene su representación estructurada.

El resultado de la actuación de TURBIO es una base de conocimiento (BCD) que contiene ejemplares de entidades del dominio presentes en la información extraida.

La adquisición del conjunto de reglas de extracción se realiza mediante tres módulos del sistema: S-BUILDER, P-BUILDER y KERNEL.

S-BUILDER tiene como objetivo crear esquemas de patrones sintáctico-semánticos relevantes para el dominio. Un esquema significa una representación del conjunto de segmentos o subsegmentos de frases que tienen el mismo árbol sintáctico-semántico. La dificultad de este proceso viene dada por el hecho de que pueden existir segmentos con una estructura sintáctica poco frecuente que contengan información relevante. En el ejemplo del Apéndice se observa que la frase "su sombrero verde pasa a blanco harinoso o ceniza blancuzco" queda analizada superficialmente en 5 segmentos. El primero de ellos, pese a no tener una estructura sintáctica frecuente en el corpus contiene información muy relevante en su subsegmento **gnom**. S-BUILDER reduce el conjunto de estructuras sintácticas relevantes en dos fases: a) elimina aquellas estructuras de segmentos y subsegmentos con baja frecuencia de aparición en el corpus (en el ejemplo estructuras 1 y 2) y b) elimina aquellas estructuras que aparecen siempre dentro de otras de alta frecuencia (en el ejemplo estructura 8).

Una vez encontradas las estructuras sintácticas relevantes, S-BUILDER produce el conjunto de esquemas de patrones sintáctico-semánticos relevantes utilizando el módulo GENERALIZER (descrito en el apartado siguiente) para encontrar la semántica de las variables etiquetadas como nombre, adjetivo o verbo. En el ejemplo se muestran los patrones sintáctico-semánticos derivados de la estructura sintáctica (sp ((r0p X) (gnom ((n Y) (a Z)))).

El módulo P-BUILDER tiene como objetivo construir patrones sintáctico-semánticos relevantes simples y compuestos. Los primeros se obtienen a partir de los esquemas intentando fijar variables de un esquema con información común de los segmentos que representa (en el ejemplo la estructura 7 da lugar al patrón (sp (r0p a) (gnom ((n [ISA:COLOR]-X) (a [ISA:MATIZ]-Y))). Los patrones compuestos son construidos concatenando patrones simples que coaparecen frecuentemente en el corpus.

Finalmente, KERNEL construye el conjunto de reglas de extracción. Cada patrón obtenido por P-BUILDER genera una regla con una prioridad de ejecución. Cuanto más largo sea el patrón mayor será la prioridad de la regla. Esto permite que patrones contenidos en otros sean evaluados posteriormente para no perder información. Las reglas llevan asociadas un *método de extracción* adecuado al patrón. Actualmente existen 5 métodos de extracción que corresponden a cinco posibles tipos de patrones, según la información relevante que contengan para rellenar tripletas <entidad, atributo, valor> de la ontología, es decir, tipos <entidad>, <entidad,valor>, <entidad,atributo,valor>, <atributo, valor> y <valor>. Las referencias y elipsis de atributos y entidades son resueltas por los métodos de extracción.

La ejecución de la extracción de información se realiza en el módulo KERNEL aplicando las reglas de extracción adquiridas sobre los segmentos analizados de todos los documentos.

3.3 Generalizador de patrones específicos

Los segmentos hallados pueden verse como patrones sintácticos. Cada constituyente del patrón trata de reconocer palabras concretas (p.e., pie, blanco), y a su vez las propiedades sintácticas de dichas palabras (p.e., N(ombre), A(djetivo)).

Si observamos el conjunto de patrones específicos, notamos que dos o más patrones poseedores de las mismas restricciones sintácticas, a pesar de compartir algunas propiedades semánticas, se aplican sobre distintas palabras. Ello sugiere que los patrones podrían ser generalizados usando su semántica.

Esta asunción supone que disponemos del conocimiento necesario sobre nuestro lenguaje. Nuestra aproximación [5] hace uso de la WordNet del castellano[7], extendida para abarcar la presentación conceptual del vocabulario del dominio de la aplicación, como conocimiento semántico de base porque su representación como red semántica nos permite razonar sobre distintos tipos de relaciones semánticas entre conceptos.

La generalización de dos patrones específicos consiste en buscar un concepto general, ascendiendo en la jerarquía semántica, que cubra ambos conceptos; es decir, hallar un concepto general para cada uno de los constituyentes. Es preciso que todos los constituyentes puedan ser generalizados en algún nivel de la jerarquía; de lo contrario, los patrones específicos no podrán ser generalizados y se mantendrán en su forma inicial.

Puesto que cada concepto específico puede tener diversas generalizaciones usando distintas relaciones, necesitamos un método que reduzca el problema. La búsqueda del concepto antecesor al ascender en la jerarquía será guiado por distintas reglas dependiendo de las relaciones presentes en la representación semántica del dominio.

[7] Forma parte de EuroWordNet (http://www.let.uva.nl/”ewn).

Una forma sencilla de hallar antecesores la proporciona la relación de hiperonimia, pero existen también otras relaciones, como la meronimia, que permiten la generalización siempre que exista algún mecanismo de control.

El siguiente ejemplo muestra, parcialmente, el proceso de generalización aplicado a un conjunto de patrones sintácticos obtenidos en el módulo precedente.

(gnom ((n "crema") (a "amarillento")))
(gnom ((n "gris") (a "ferruginoso")))
(gnom ((n "pie") (a "blanco")))
(gnom ((n "blanco") (a "harinoso")))
(gnom ((n "ceniza") (a "blancuzco")))
(gnom ((n "sombrero") (a "rosado")))

Supongamos que la lista de segmentos anterior, es una lista completa de todos los segmentos obtenidos a partir del documento de validación. Los dos primeros segmentos no permiten ningún tipo de generalización puesto que la semántica de "crema" (que es un matiz) no posee ningún concepto en común a ningún nivel de la jerarquía con la semántica de "gris" (que es un color). Es justo el comportamiento deseado: una propiedad de un color no es un color.

Entre los tres primeros segmentos no es posible generalización alguna por razones similares. Cuando pasamos a considerar el cuarto segmento si es posible una generalización entre los segmentos primero y cuarto. La semántica de "gris" (que es un color) y la semántica de "blanco" (que es un color) tienen un concepto en común en la jerarquía y, a su vez, la semántica de "ferruginoso" (que es un matiz) y la de "harinoso" (que es un matiz) tienen un concepto en común. La representación del patrón generalizado resultante es la siguiente:

(gnom ((n [ISA:COLOR]-X) (a [ISA:MATIZ]-Y)))

Siguiendo el ejemplo también es posible una generalización entre los segmentos primero y quinto, y entre el tercero y el sexto dando lugar a:

(gnom ((n [ISA:MATIZ]-X) (a [ISA:COLOR]-Y)))
(gnom ((n [ISPARTOF:SETA]-X) (a [ISA:COLOR]-Y)))

Es interesante hacer notar que no hay pérdida de información en el proceso de generalización. El conjunto de patrones generalizados cubren totalmente el conjunto de segmentos inicial y, aún más importante, cubre nuevos ejemplos similares y reduce la cantidad de patrones específicos.

El proceso de generalización es un proceso de realimentación. Se aplican distintas reglas en distintos ejemplos para poder obtener una generalización y precisamente esta generalización puede resultar demasiado general para el tipo de información que se desea extraer. Necesitaremos un experto en el dominio para que detecte la sobregeneralización y limite la aplicación de las reglas de generalización. Así, el proceso de generalización se repetirá bajo nuevas restricciones evitando las reglas causantes de la sobregeneralización.

4. Sistemas multilingües de extracción de información

Los SEI serán tanto más útiles cuanto más versátiles sean. En este sentido son muy interesantes los trabajos encaminados a diseñar SEI multilingües, es decir, sistemas en los cuales los documentos almacenados están en diferentes lenguas y que aceptan consultas también en diferentes lenguas.

Al igual que para diseñar un SRI, un SEI multilingüe puede plantearse de maneras diferentes: usar técnicas de traducción automática, basarse en thesaurus, y usar corpus [6].

El problema de los actuales sistemas de traducción automática es que sólo son capaces de producir traducciones de alta calidad en dominios restringidos.

Tanto en un SRI como en un SEI, la precisión semántica es más importante que el análisis sintáctico y por tanto es necesario un amplio conocimiento del dominio.

Hay que tener en cuenta que podemos decidir traducir los documentos (para tenerlos todos almacenados en la misma lengua), las consultas (quizás incluso a varias lenguas si los documentos son heterogéneos) o ambos. Siempre resultará más difícil la traducción de una consulta que la de un documento pues el contexto que puede manejar el analizador es más restringido.

Un planteamiento diferente consiste en utilizar un thesaurus multilingüe. Un thesaurus es una herramienta que permite organizar la terminología propia de un dominio, es decir, es una ontología especialmente diseñada para organizar terminología.

Los thesaurus permiten asociar términos y conceptos de forma comprensible para las personas. Los multilingües permiten conectar términos en diferentes lenguas con el mismo concepto. Así pues, la extracción de información se basaría en localizar el concepto adecuado a partir de los términos expresados en cualquiera de las lenguas que el thesaurus tenga previstas. Un thesaurus multilingüe codifica varios aspectos del conocimiento del dominio: la sinonimia entre lenguas, las relaciones jerárquicas entre conceptos y las relaciones asociativas. El thesaurus debe ser lo mas amplio posible a nivel terminológico y a nivel conceptual. En general el coste de esta tarea ha llevado a los investigadores a explorar técnicas basadas en corpus (métodos estadísticos o híbridos o métodos de aprendizaje automático).

Los SEIs actuales no suelen tener en cuenta el aspecto multilingüe. En el marco del proyecto ITEM ya citado, pretendemos aplicar las técnicas aquí descritas a la EI multilingüe (las lenguas implicadas son el castellano, el catalán y el vasco). El soporte básico para esta extensión lo constituye una ontología léxica multilingüe actualmente en construcción[8].

5.- Conclusiones

La extracción de información se revela como una de las aplicaciones del TL con mayor proyección. En este artículo hemos presentado un entorno de extracción de información de textos de dominio restringido. El entorno integra diferentes módulos que abordan distintos problemas del TL (análisis léxico y morfológico, análisis sintáctico parcial, aprendizaje de reglas de extracción, extracción de información, etc...). El módulo de extracción ha sido aplicado con éxito en el dominio micológico. Actualmente se trabaja en su extensión a un sistema de extracción multilingüe.

Bibliografía

[1] J. Atserias, H. Rodríguez. *TACAT: TAgged Corpus Analizer Tool.* Report interno LSI-RT-2-98.

[2] J. Carmona, L. Márquez, L. Padró, H. Rodríguez, J. Turmo. *An Environment for Morphosintactic Processing of Unrestricted Spanish Text.* Actas de First internation Conference of Language Resource and Evaluation. Granada, 1998.

[3] L. Márquez, L. Padró. *A Flexible POS Tagger Using an Automatically Acquired Language Model.* Actas de ACL-COLING'97. Madrid, 1997.

[4] J. Turmo, N. Català, H. Rodríguez. *TURBIO: A System for Extracting Information from Restricted-domain Texts.* Actas de IEA-AIE'98. Castellón, 1998.

[5] N. Català, N. Castell. *Construcción Automática de Diccionarios de Patrones de Extracción de Información*. Actas de SEPLN'97. Madrid, 1997.

[6] C. Fluhr. *Multilingual information retrieval.* En R.A. Cole et al. editores. Survey of the State of the Art in Human Language Technology. Center for Spoken Language Understanding.Oregon,1995. htttp://www.cse.ogi.edi/CSLU/HLTsurvey/ch8node7.html

---

[8] El modelo de esta ontología es EuroWordNet ,ya citado.

Apéndice

S-BUILDER

Frase: "Su sombrero verde pasa a blanco harinoso o ceniza blancuzco".

Análisis superficial: (sn ((r0a su) (gnom ((n sombrero) (a verde))))) (v0v pasa)
(sp ((r0p a) (gnom ((n blanco) (a harinoso)))) (c0c o)
(gnom ((n ceniza) (a blancuzco))))

Generalización a estructuras sintácticas:

	frecuencia en corpus	estructuras sintácticas
1	5	(sn ((r0a X) (gnom ((n Y) (a Z))))
2	5	(r0a X)
3	105	(gnom ((n X) (a Y)))
4	230	(n X)
5	519	(a X)
6	309	(v0v X)
7	195	(sp ((r0p X) (gnom ((n Y) (a Z))))
8	195	(r0p X)
9	931	(c0c X)

Reducción de esquemas:

1 y 2: no relevantes por baja frecuencia
4: cubierto parcialmente por 3
5: cubierto parcialmente por 3 y 7 (1 ha sido eliminado)
8: cubierto por 7 totalmente

GENERALIZER

Generalización a esquemas sintáctico-semánticos:

(gnom ((n [ISA:COLOR]-X) (a [ISA:MATIZ]-Y)))

(gnom ((n [ISPARTOF:SETA]-X) (a [ISA:COLOR]-Y)))

(gnom ((n [ISA:MATIZ]-X) (a [ISA:COLOR]-Y)))

P-BUILDER

Obtención de patrones sintáctico-semánticos:

(sp ((r0p X) (gnom ((n [ISA:COLOR]-Y) (a [ISA:MATIZ]-Z)))

X = (palabra: ‘a’, frecuencia:195)

$\Rightarrow$ (sp ((r0p a) (gnom ((n [ISA:COLOR]-Y) (a [ISA:MATIZ]-Z)))))