Ensayo de un sistema de extracción de información (técnica de inteligencia artificial) en un centro de información especializado en sanidad vegetal.

Por:
Ingrid Paz García
Ejecutivo de Informática y Comunicaciones. Dirección Nacional Gaviota S.A. (Cuba)
Correo electrónico: ingrid@gaviota.gav.tur.cu

Resumen

Se confeccionaron automáticamente reglas que sirven para el proceso de extraer información de textos con resúmenes de artículos de información científica acerca de Control de Plagas de Plantas del Instituto de Investigaciones de Sanidad Vegetal. Para el desarrollo de la misma se tomó como punto de partida el sistema CRYSTAL, que induce automáticamente a través de ejemplos un conjunto de reglas para el análisis de textos en dominios específicos. El sistema desarrollado para lograr el objetivo propuesto fue denominado SEISAV (Sistema de Extracción de Información en Sanidad Vegetal).

Palabras claves:
Sistemas expertos, Inteligencia artificial, Procesamiento del lenguaje natural, Sanidad vegetal

Abstract

Rules for process of extracting text information with abstracts of scientific information articles about plants pests control from Plan Health Research Institute were carried out automatically. For it's develop, CRYSTAL system was taken as departure point. This system induces automatically throughout examples, rules gadgetry for text analysis in specific domains. The system developed to achieve the objective proposed was denominated SEISAV (Information Extraction System on Plant Health)

Key words:
Expert systems, Information extraction system, Artificial intelligence, Natural language processing, Plant health

Introducción

La Inteligencia Artificial se ha definido por varios autores como: diseño de sistemas inteligentes que exhiben características que se asocian con la inteligencia humana -entender lenguaje natural, aprendizaje, razonamiento, etc. (Feigenbaum) o campo de la ciencia y de la ingeniería que se ocupa de la comprensión a través de la computadora de lo que comúnmente se llama comportamiento inteligente y de la creación de herramientas que exhiben tal comportamiento (Shapiro). Dentro de sus múltiples aplicaciones se encuentra la comprensión del lenguaje natural. (Morales, 1999).

En la actualidad, la información constituye un elemento de gran importancia dentro de nuestras vidas. La tecnología moderna nos ha puesto delante volúmenes increíbles de información, mucha de la cual, por estar disponible en textos escritos en lenguaje natural sin restricciones, necesita de un procesamiento previo para poder ser usada y aplicada a la resolución de los problemas que tenemos que enfrentar.

Debido al gran almacenamiento de información existente en textos, el tiempo que requiere su procesamiento manual para la extracción de aquella que resulta relevante es muy pequeño. Por eso se va haciendo imprescindible el uso de sistemas automáticos que ayuden a procesar o extraer el contenido conceptual encerrado en esos volúmenes. Un aspecto fundamental de este proceso de información tiene que ver con la cantidad de conocimiento que deberá ser aplicado para la extracción de información relevante a la solución del problema, el cual, por lo general, sólo lo tienen los expertos en el problema y la necesidad, por tanto, de hacer asequible y manipulable este conocimiento para su resolución.

Una de las estrategias más comúnmente adoptadas es la Recuperación de Información, pero la Extracción de Información –técnica de Inteligencia Artificial- es una estrategia diferente, ya que a partir de la primera se obtienen documentos con información significativa, mientras que un Sistema de Extracción obtiene hechos de los documentos, o sea, extrae y organiza la información relevante e ignora la irrelevante. Las dos técnicas, son, por tanto, complementarias. (University of Sheffield, 2002).

Los Sistemas de Extracción de Información (SEI) operan en un contexto formado por un conjunto de textos en lenguaje natural para extraer determinados conceptos que son de nuestro interés para una aplicación específica. Estos textos, en unión de la información definida para ser extraída, conforman el *dominio* de trabajo de un SEI.

Aunque la idea de extraer automáticamente información escrita en lenguaje natural data de principios de los años 80, el campo de la extracción de información fue creado a finales de esa década por DARPA (Defense Advanced Research Projects Agency), enfocando este proceso en gran medida a una tarea específica [Appelt and Israel, 1999]. DARPA instituyó por esta época las Conferencias o Competiciones para el entendimiento de mensajes, conocidas como **MUCs**, en respuesta a las oportunidades representadas por la enorme cantidad de textos en formato electrónico disponibles.

En cada conferencia se presenta un dominio sobre el cual los sistemas compiten para ver cuál logra los mejores resultados, según las especificaciones impuestas en la tarea de extracción de información. Los dominios presentados en los MUCs hasta 1997 han sido: textos sobre operaciones navales [**MUC-1**, 1987 y **MUC-2**, 1989], noticias sobre actividades terroristas [**MUC-3**, 1991 y **MUC-4**, 1992], noticias sobre microelectrónica y fusión de corporaciones [**MUC-5**, 1993], artículos sobre sucesión de puestos en compañías importantes [**MUC-6**, 1995], artículos sobre vehículos espaciales y lanzamiento de misiles [**MUC-7**, 1997]. [Appelt and Israel, 1999]

Producto de los MUCs han sido confeccionados varios sistemas, por ejemplo: **Autoslog** [Riloff, 1993], **PALKA** [Kim y Moldovan, 1995], **LIEP** [Huffman, 1995], **HASTEN** [Krupka, 1995], citados por Glickman y Jones, [1999] y Muslea [1999]; **CRYSTAL** [Soderland et al, 1995], **FASTUS** [Hobbs *et al*, 1995], **AutoSlog-TS** [Riloff y Shoen, 1995; cit. Glickman y Jones, 1999] que es una extensión de Autoslog. Además, **RESOLVE** [McCarthy y Lehnert, 1995], **MLR** [Aone y Bennett, 1995], **RAPIER** [Califf y Mooney, 1997], **Nymbl**e(BBN) [Bikel *et al*, 1997], **MEN**E(NYU) [Borthwick *et al*, 1998], también referidos por Glickman y Jones [1999].

Un SEI necesita consolidar la salida proporcionada por la aplicación de estas reglas específicas usando un paso posterior denominado *procesamiento o análisis de discurso.* Este implica tres problemas:

- el análisis de frases nominales
- la resolución de co-referencias (determina cuándo se refiere a una misma cosa descripciones diferentes y es sensible a las características estructurales del texto, a las semánticas y al correcto análisis de las oraciones y de las frases nominales complejas)
- el reconocimiento de enlaces relacionales

Otra cuestión importante radica en considerar las reglas como dependientes o independientes del dominio, pues una cosa es crear una base de heurísticas para una aplicación específica y otra, crear capacidades de análisis independientes del dominio.

La filosofía de trabajo de los SEI se basa en la aplicación de un conjunto de reglas construidas, tanto manual como automáticamente, para identificar las referencias a la información que nos interesa dentro de una serie de textos y proporcionar una representación simbólica de la misma. *Estas reglas están basadas en aspectos del vocabulario, de la semántica y del estilo de escritura propios de cada dominio* [Soderland, 1997]. Por esto la utilización de técnicas para el procesamiento del lenguaje natural y el uso de conocimiento relacionado con el dominio en que se está trabajando son de vital importancia para la construcción de un SEI.

Desarrollo

El trabajo se enfocó en lograr la confección automática de reglas que sirvan para el proceso de extraer información de textos con resúmenes de artículos de información científica acerca de Control de Plagas de Plantas del Instituto de Investigaciones de Sanidad Vegetal. Para su desarrollo se tomó como punto de partida el sistema CRYSTAL [Soderland *et al*, 1995], que induce automáticamente a través de ejemplos un conjunto de reglas para el análisis de textos en dominios específicos. El sistema desarrollado para lograr el objetivo propuesto fue denominado SEISAV (Sistema de Extracción de Información en Sanidad Vegetal).

El trabajo en el dominio de Control de Plagas de Plantas se define como el análisis de resúmenes de artículos de investigación científica desarrollada en esta materia para identificar la información concerniente a las plagas que se sometieron a control, el control ejercido y los cultivos en los cuales se aplicó dicho control. La siguiente oración extraída de un artículo, ilustra el tipo de información a manipular: *La aplicación de Bacillus thuringiensis contra Spodoptera frugiperda incrementó el rendimiento del cultivo del maíz.*

Esta oración contiene la siguiente información: *Bacillus thuringiensis* se usó para controlar la plaga *Spodoptera frugiperda* en el cultivo del maíz. Esta puede ser representada de manera simbólica utilizando un **record** con tres campos:

- **Control(es)**: Bacillus thuringiensis
- **Plaga(s)**: Spodoptera frugiperda

- **Cultivo(s):** maíz

Los campos estarían destinados a almacenar los conceptos a extraer en este dominio: plagas, cultivos y tipos de control.

¿Cómo es posible que un SEI reconozca dentro de un texto que la oración del ejemplo posee información relevante y sea capaz de extraer esta información representándola como se muestra en el record con tres campos? Esto se logra aplicando una serie de niveles o etapas de procesamiento. El primer nivel consta de un análisis sintáctico que identifica la oración y sus principales constituyentes sintácticos. Estos constituyentes son: el *Sujeto*, la *Forma Verbal* y los *Complementos*.

El segundo nivel asocia cada palabra dentro de cada constituyente, si es posible, con conceptos semánticos propios del dominio. Para este dominio fue creado, por especialistas en la materia, un conjunto de conceptos semánticos vinculados entre sí por relaciones de subclasificación para poder clasificar los términos que tuvieran significación para el dominio. Entre estos conceptos se encuentran el de "**Control**", que abarca todos los tipos de controles de plagas existentes y el de "**Cultivos Permanentes**", que representa los términos que se refieren a estos tipos de cultivos.

En el tercer y último nivel se aplican un conjunto de reglas de análisis de textos específicas para el dominio, con el fin de encontrar las referencias a la información que se desea extraer.

La función de estas reglas consiste en buscar un conjunto de determinadas evidencias lingüísticas dentro de los textos que respondan a estructuras semánticas y léxicas, como las que se muestran a continuación:

1. <**Plagas**> encontrar incidir en <**Cultivos**>
2. experimentar <**Cultivos**>
3. <**Control**> aplicar en <**Cultivos**>

Si estas evidencias son encontradas dentro de una oración, se puede decir que en ésta se encuentra información que resulta relevante para el dominio y entonces, con los conceptos de interés identificados en ella, se crearía un record para representar esta información, el cual se ofrecería como salida a la aplicación de dicha regla.

Las evidencias lingüísticas pueden ser conceptos semánticos propios del dominio y determinadas palabras claves. Por ejemplo, una regla que busque evidencias lingüísticas que se ajusten al segundo conjunto de estructuras semánticas y léxicas del ejemplo, identificaría la oración "Se experimentó en los cultivos de tabaco de la provincia de Pinar del Río" como una oración con información de interés, ya que ésta presenta evidencia del verbo "experimentar" en las palabras "se experimentó" y evidencia del concepto semántico "**Cultivos**" en el término "tabaco".

El dominio de Control de Plagas de Plantas consta de un conjunto de textos seleccionados de libros de resúmenes del Forum de Manejo Integrado de Plagas [Resúmenes de Ponencias. Forum Tecnológico sobre Manejo Integrado de Plagas, 1998] y de resúmenes de artículos publicados en la revista *Fitosanidad* del Instituto de Investigaciones de Sanidad Vegetal [INISAV, 1998].

La información a ser extraída concierne solamente a las plagas que se sometieron a control, al control que se aplicó y a los cultivos en donde se experimentó. Cualquier otra información como: plagas que no se controlaron, controles no efectivos o no aplicados y cultivos que no se tuvieron en cuenta, es considerada como irrelevante y por lo tanto, ignorada.

No siempre se puede juzgar la relevancia de una información basada en contexto local [Soderland, 1997]. Supongamos que se analiza la siguiente oración: "El experimento se realizó en el cultivo de la papa". La pregunta sería si presenta información relevante a tener en cuenta para este dominio. Si el experimento realizado fue un control de plagas de plantas, entonces se puede decir que esta oración es portadora de información interesante, en este caso sería la concerniente al cultivo en el cual se experimentó. Si el experimento hubiera sido en otro tema, por ejemplo, la incidencia de fertilizantes sobre el rendimiento de los cultivos, entonces esta información no hubiera sido de ninguna relevancia. *Las reglas de extracción de información que están basadas en un contexto local no pueden hacer tales diferenciaciones y deben estar sujetas a un posterior análisis para eliminar las extracciones incorrectas* [Soderland 1997].

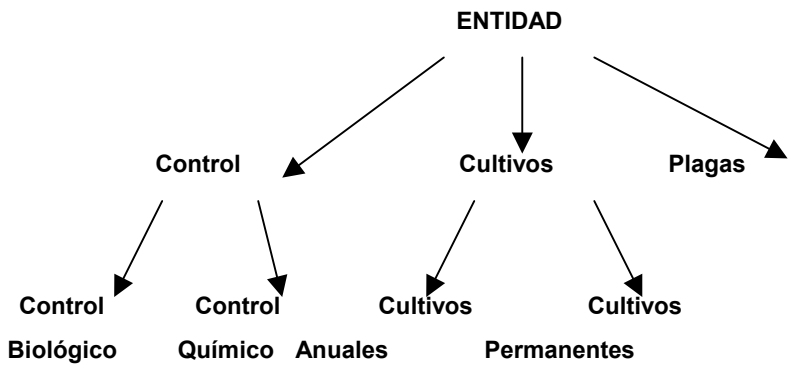
La salida definida para este dominio es representada como un record con tres campos: **Plaga(s)**, **Control(es)** y **Cultivo(s)**, los cuales no tienen que estar necesariamente ocupados. Una oración que contiene información relevante puede llenar completamente el record o dejar hasta dos campos vacíos. Es importante destacar que en un campo las extracciones pueden ser múltiples.

Debido a esto, las reglas construidas no tienen que buscar necesariamente por evidencias semánticas que representen a cada uno de los conceptos a tener en cuenta para extraer información relevante para el dominio, así como cada evidencia semántica puede identificar a más de un término dentro de los constituyentes sintácticos de una oración.

Con el objetivo de representar los conceptos semánticos en un dominio y las relaciones entre éstos, los SEI operan con una jerarquía de clases semánticas. Cada clase representa a un concepto semántico y la posición de estas clases dentro de la jerarquía está determinada por las relaciones existentes entre los conceptos semánticos que representan.

Esta jerarquía nos brinda información sobre cuáles conceptos son los más específicos y cuáles los más generales para clasificar a determinados términos propios del dominio. Por ejemplo, la jerarquía de clases semánticas diseñada para SEISAV representa a conceptos como "**Control Botánico**" con la clase "**Control_Botánico**" y "**Control Zoológico**" con la clase "**Control_Zoológico**" y expresa de forma simbólica que el concepto "**Plagas**" nos ofrece una definición más general acerca de los organismos dañinos para las plantas que la que brindan los conceptos "**Organismos Fitófagos**" y "**Organismos Fitopatógenos**".

Los conceptos semánticos propios del dominio se comenzaron a representar a partir de las clases "**Control**", "**Cultivos**" y "**Plagas**". La raíz de la jerarquía, la clase "**ENTIDAD**", no representa ninguna información semántica a tener en cuenta.



Las clases que conforman esta jerarquía se dividieron en **clases principales** y **clases modificadoras**. Las clases principales son las que representan los conceptos semánticos más específicos y las modificadoras son las que representan los conceptos semánticos con un grado mayor de generalidad. SEISAV utiliza clases principales, como la que representa al concepto "**Bacterias Parásitos**", ya que

este concepto define al término "Bacillus thuringiensis*"* con la mayor especificidad posible, mientras que la clase que representa al concepto "**Microorganismos Parásitos",** que define a este término más generalmente, se considera como clase modificadora.

Otra utilidad de esta jerarquía es que permite a una regla que busca la presencia de determinados conceptos semánticos en una oración, encontrar no sólo aquellos términos que se pueden clasificar directamente con estos conceptos, sino también aquellos términos que puedan clasificarse usando conceptos semánticos que mantengan una relación de subclasificación con los conceptos por los que busca dicha regla. Por ejemplo, una regla que busque la presencia del concepto semántico "**Control**" en el sujeto podría ser aplicada con resultados positivos a oraciones en cuyos sujetos aparezcan términos que pueden clasificarse usando los conceptos semánticos "**Control Biológico**" y "**Control Químico**", ya que estos son abarcados por el concepto "**Control**".

I. Construcción de reglas de análisis de texto para un dominio específico

Las reglas para el análisis de textos construidas para un dominio no pueden, en general, ser utilizadas en un dominio diferente [Soderland 1997]. Esto es debido a que durante el proceso de desarrollo de estas reglas no sólo se utiliza información relacionada con el lenguaje, también se usa contenido semántico que es inherente a cada dominio en particular y la información que se desea extraer no es igual en cada dominio. Por ejemplo, en el dominio en que estamos trabajando las reglas para extraer información no serían de ninguna utilidad en un dominio médico, en el cual los conceptos de interés serían diagnósticos, padecimientos, etc. de las historias clínicas de los pacientes en un hospital. Las reglas para el dominio de Control de Plagas de Plantas están relacionadas con información que abarca los temas agronómico y biológico.

Para construir estas reglas se pueden usar dos métodos: el método de "Ingeniería por Conocimiento" o manual y el método de "Entrenamiento automático" [Appelt e Israel, 1999]. En el método manual las reglas son construidas por una persona familiarizada con los SEI y con el formalismo de expresar reglas de análisis de textos para estos sistemas, la cual debe contar con la ayuda de un experto en el dominio. El método de "Entrenamiento automático" es completamente diferente. Este no requiere tener a mano a alguien con conocimiento de cómo un sistema de extracción de información opera o de cómo construir reglas para éste. Sólo es necesario contar con un experto en la materia y con textos propios del dominio que sirvan de materia prima en la confección de las reglas. *Esto se debe a que las reglas son creadas automáticamente a través de un aprendizaje que involucra a los textos: éstos son previamente* ***anotados*** *por el experto en el dominio con el objetivo de proporcionar la información relevante como ejemplos para que el sistema disponga de una guía para construir las reglas* [Soderland, 1997].

El concepto de **anotar** se define como marcar dentro de los textos los términos que poseen algún significado semántico para el dominio en que se opera.

Estos métodos presentan sus ventajas y desventajas. *Las desventajas del método manual radican en que no siempre se puede contar con la requerida presencia de un profesional en la materia, en que este proceso resulta bastante trabajoso y que un cambio en las especificaciones del problema puede ser difícil de acomodar. La principal ventaja que ofrece este método es que las reglas construidas manualmente logran un alto rendimiento en el proceso de extraer información. Por su parte, las desventajas del método de "entrenamiento automático" radican en el gran volumen de textos con que hay que contar para el entrenamiento, en que estos textos pueden ser difíciles de conseguir, y en que un ligero cambio en las especificaciones de la información a extraer conllevaría volver a anotar una gran cantidad de textos. La ventaja de éste se refiere a la facilidad con que el sistema construido*

usando este método para la confección de las reglas puede adaptarse a nuevos dominios [Appelt and Israel 1999].
No se pueden hacer consideraciones sobre cuál método es superior ya que existen situaciones en las que uno aventaja a otro y viceversa. Un aspecto que puede decidir es que los rendimientos más elevados se han obtenido usando reglas construidas manualmente. *Pero las reglas diseñadas automáticamente brindan la posibilidad de descubrir patrones que no hayan sido vistos por un experto y están construidas con un grado justo de generalización* [Appelt e Israel, 1999]. En la actualidad los sistemas que más abundan operan usando reglas obtenidas mediante métodos automáticos.

De acuerdo con lo anterior, consideramos que esta tendencia a utilizar el método automático radica en la ventaja anteriormente mencionada: la facilidad con que estos sistemas se adaptan a nuevos dominios. *Los sistemas que inducen automáticamente las reglas de análisis de texto pueden ser modificados para trabajar en otros dominios sin la participación de una persona familiarizada con la filosofía de trabajo de los SEI. Sólo se requeriría la presencia de un experto en el nuevo dominio que realizara la anotación de los textos de ejemplos y que especificara los conceptos semánticos a tener en cuenta* [Appelt e Israel 1999].

II. Rendimiento

Cuando hablamos de rendimiento de un sistema de extracción de información nos estamos refiriendo a dos parámetros que nos dan una medida de éste. El rendimiento nos dice cuán eficientemente se comportan las reglas de análisis de texto extrayendo información relevante e ignorando aquella que no interesa en el dominio.

Estos parámetros son **precisión** y **recobrado**. La **precisión** es el porcentaje de extracciones realizadas que fueron correctas y el **recobrado** es el porcentaje de oraciones con información relevante para el concepto a extraer que fueron correctamente identificadas. Esto se traduce en: precisión es igual a la relación entre el número de oraciones con información relevante de las que se extrajeron correctamente los conceptos definidos y la suma de éstas con el número de oraciones con información irrelevante que el sistema consideró que poseían conceptos de interés; y recobrado es igual a la relación entre el número de oraciones con información relevante de las que se extrajeron correctamente los conceptos definidos y la suma de éstas con el número de oraciones que contenían información relevante que fueron consideradas por el sistema como oraciones con información irrelevante. A continuación se ofrece una definición más formal de estos parámetros:

Precisión = **PE**/(**PE** + **NE**)

Recobrado = **PE**/(**PE** + **PNE**)

PE: Oraciones con información relevante extraída

NE: Oraciones con información irrelevante extraída

PNE: Oraciones con información relevante no extraída

Los valores de estos parámetros se obtienen aplicando las reglas construidas sobre un conjunto de textos propios del dominio que no fueron usados durante el entrenamiento. Este conjunto de textos se denomina **conjunto de prueba no visto**.

Para el caso de medir el rendimiento de SEISAV sólo se tuvo en cuenta el parámetro recobrado, ya que los resúmenes de artículos científicos del tema de Control de Plagas de Plantas presentaban solamente información considerada como positiva en este dominio. Generalmente, para no ser absolutos, los artículos de corte científico en esta materia sólo publican aquellos resultados que fueron satisfactorios, por lo que no estaban disponibles textos con información irrelevante o no interesante.

III. Anotación de los textos

Para poder inducir automáticamente el conjunto de reglas que serán usadas para la extracción de contenido conceptual, se necesita de textos en los que se encuentren identificados los términos que representan el tipo de información a tener en cuenta para el dominio. Cada oración, en el conjunto de textos de entrenamiento que se usó para la confección de estas reglas, contenía información relevante y todas fueron explícitamente anotadas con el objetivo de identificar los términos que representaban esta información. Estas oraciones anotadas se pueden definir como **instancias**.

El proceso de anotar es simple. Se marcan los términos que representan contenido semántico, es decir, los términos que puedan clasificarse usando los conceptos semánticos definidos para el dominio. Anotar o marcar un término que nos interesa en una oración significa incluir etiquetas que lo identifiquen al inicio y al final de éste. Por ejemplo, para anotar un término que representa al concepto "**Control Botánico**" se usa la etiqueta <Cont_Bot> para marcar el inicio y </Cont_Bot> para marcar el final. "Cont_Bot" es la abreviatura de la clase semántica "**Control_Botánico**", que es la que representa en la jerarquía al concepto semántico *"***Control Botánico**". A continuación se muestra una oración completamente anotada:

> La aplicación de <Bact_Par>Bacillus thuringiensis</Bact_Par> contra <Ins_Fit>Spodoptera frugiperda</Ins_Fit> incrementó el rendimiento del <Cult>cultivo</Cult> de <Cult_Anual>maíz </Cult_Anual>.

Producto de este proceso de anotación se va obteniendo un fichero de texto en el cual se almacenan las oraciones que fueron marcadas, y varios ficheros de texto, cada uno representando a un concepto de los especificados para el dominio, con el fin de ir confeccionando diccionarios técnicos con los términos que son marcados.

Con el objetivo de agilizar el proceso de medir el rendimiento de las reglas de análisis de texto inducidas, las oraciones que formaban parte del conjunto de prueba no visto también fueron anotadas. En realidad, a la hora de anotar no se especificó cuáles oraciones serían usadas en la fase de entrenamiento y cuáles en la fase de prueba, debido a que se realizaron varios experimentos para comprobar el efecto de las dimensiones del texto de entrenamiento sobre el rendimiento del sistema

IV. Reglas para el análisis de textos de SEISAV

SEISAV es un sistema que construye automáticamente reglas para extraer conceptos de interés en un dominio particular. Estas reglas están formadas por estructuras sintácticas dentro de las cuales se encuentran combinaciones de conceptos semánticos propios del dominio y términos léxicos a los que llamaremos **restricciones**. Para poder identificar cierto concepto de interés para el dominio dentro de un texto, las restricciones de una regla deben ser cumplidas por determinadas evidencias lingüísticas de algún fragmento del texto que, para el caso de SEISAV, va a constituir una oración. Las reglas se

representaron por un récord que denominaremos **Definición Conceptual**, el cual describiremos con más detalle posteriormente.

Es importante que las reglas tengan acceso a cierto conocimiento sintáctico, al menos, el conocimiento de distinguir los principales constituyentes sintácticos que conforman una oración [Soderland, 1997]. Para lograr construir reglas formadas por estructuras sintácticas propias de la gramática española se diseñó e implementó un analizador sintáctico que fuera capaz de dividir una oración en sus principales constituyentes sintácticos. Estos constituyentes son el sujeto, la forma verbal y los complementos.

Los conceptos semánticos son también de gran importancia. *Expresar las reglas usando conceptos semánticos posibilita reglas compactas con un grado de generalización mucho mayor que usando solamente términos léxicos* [Soderland, 1997]. Por ejemplo, supongamos que existe una regla que posea una restricción en el sujeto formada por el concepto semántico "**Control Microbiológico**". Dicha restricción será satisfactoriamente cumplida por oraciones en las que en el sujeto aparezcan los términos "Trichoderma harzianum", "Paecilomyces lilacinus", "Beauveria bassiana" y todos aquellos que se puedan clasificar bajo el concepto semántico "**Control Microbiológico**" que, para este ejemplo, no constituye el concepto semántico que con más especificidad define a dichos términos. Esto último, el hecho de poder usar un concepto más general para identificar evidencias semánticas que se clasifican por conceptos más específicos, constituye otra facilidad de la utilización de conceptos semánticos para expresar las reglas. El uso de conceptos semánticos posibilita que una restricción pueda ser satisfactoriamente aplicada a una gran cantidad de términos dentro de una oración. La jerarquía de clases semánticas que se usó para la representación de los conceptos semánticos del dominio, presentada en el capítulo anterior, es la que posibilita la realización de estos procesos que involucran a la semántica.

En algunos casos el uso de conceptos semánticos no es suficiente. *La utilización de términos léxicos junto a los conceptos semánticos es importante cuando estos últimos no están lo suficientemente afinados para realizar algunas distinciones necesarias para la tarea de extraer información* [Soderland, 1997]. En el dominio de Control de Plagas, por ejemplo, la palabra "aplicación" constituye una pista para evidenciar el concepto "**Control**" ("la aplicación de Bacillus thuringiensis").

V. Análisis sintáctico

El análisis sintáctico constituye la base para la construcción de las reglas de análisis de textos. Los constituyentes sintácticos, unidos a las evidencias semánticas y léxicas dentro de los mismos, son los elementos que conforman la representación de las instancias para el entrenamiento y de las reglas después de construidas.

Lo primero fue analizar la estructura sintáctica de las oraciones con información relevante que conformaban los textos seleccionados. Se determinó que lo más conveniente era dividir la oración en cinco constituyentes sintácticos: Sujeto, Forma Verbal, Complemento Directo, Complementos Indirectos y Complementos Circunstanciales. Para el Sujeto y el Complemento Directo solamente se extrajo la información de los términos que los formaban, sin hacer distinción en las funciones gramaticales de éstos. Para los Complementos Indirectos y Circunstanciales se extrajeron los términos que los formaban haciendo distinción de las preposiciones que los introducían[1]. Para la Forma Verbal se extrajo información acerca del verbo, de la forma reflexiva del pronombre personal de la tercera

[1] Las preposiciones que introducen a los complementos indirectos son ***a*** y ***para***. Los complementos circunstanciales no tienen por qué estar obligatoriamente introducidos por preposiciones.

persona usada como proclítico[2] y del modo en que se encontraba el verbo, o lo que es lo mismo, si el verbo se encontraba negando o afirmando. A continuación se muestra un esquema con los constituyentes sintácticos y la información sintáctica asociada a éstos:

SUJETO :: términos:
FORMA VERBAL :: auxiliar:
verbo:
modo:
COMPLEMENTO DIRECTO :: términos:
COMPLEMENTOS INDIRECTOS :: términos:
preposiciones:
COMPLEMENTOS CIRCUNSTANCIALES :: términos:
Preposiciones:

Posteriormente se comenzó la construcción de una gramática computacional que identificara la estructura de oraciones de la gramática española. Con esta finalidad se diseñó una *gramática libre de contexto* (GLC). Los símbolos utilizados representan a los sustantivos, a los adjetivos, a los adverbios, a las formas preposicionales, entre otros componentes sintácticos. A continuación se muestra una relación completa de estos símbolos así como la función que desempeñan dentro de la GLC (si son símbolos terminales o no terminales):

Símbolos	**Tipo**	**Componente gramatical que representa**
O	no terminal	Oración
S	no terminal	Sujeto
FV	no terminal	Forma Verbal
CD	no terminal	Complemento Directo
CI	no terminal	Complemento Indirecto
CC	no terminal	Complemento Circunstancial
FP	no terminal	Forma Preposicional
FS	no terminal	Forma Sustantiva
P/N	no terminal	Positivo/Negativo
SUST	terminal	Sustantivo
ADJET	terminal	Adjetivo
PREP	terminal	Preposición
PREPCC	terminal	Preposición complemento circunstancial
PREPCI	terminal	Preposición complemento indirecto
ADVERB	terminal	Adverbio
CONJ	terminal	Conjunción
V	terminal	Verbo
AUX	terminal	Auxiliar
ART	terminal	Artículo
NEG	terminal	Negación
COMA	terminal	Coma
0	terminal	Vacío

[2] Dícese de la palabra sin acentuación prosódica que se liga en la oración con el vocablo siguiente

Las reglas que forman la gramática están dirigidas a lograr una división sintáctica que responda a la estructura de las oraciones con información relevante presente en los textos que se usaron para el entrenamiento y la prueba del sistema.

Otra aclaración importante es la diferenciación que se realizó con las preposiciones. El símbolo "PREPCI" se utilizó para representar a las preposiciones "a" y "para", que son las que encabezan los complementos indirectos, como se especificó anteriormente. El símbolo "PREPCC" se usó para la representación de las preposiciones que pudieran encabezar a los complementos circunstanciales, ya que en las oraciones de los textos este tipo de complemento estaba encabezado fundamentalmente por ciertas preposiciones como "en", "sobre", "entre", etc.

Para cada uno de estos componentes sintácticos se construyeron diccionarios representados por ficheros de sólo texto, por ejemplo, para los sustantivos se usó el fichero "Sust.wri". Los términos abarcados por la jerarquía de clases semánticas se consideraron como sustantivos por ser nombres propios. Para estos términos, como resultado del proceso de anotación de los textos, también se construyeron diccionarios.

VI. Definiciones Conceptuales

Una **Definición Conceptual** es un récord cuyos campos están destinados a almacenar información tanto semántica como léxica. Esta información, que se encuentra relacionada con las estructuras sintácticas, puede ser usada, como se dijo anteriormente, para representar a las reglas de análisis de textos. Para esto, los campos del récord sostienen la información de las restricciones que forman a dichas reglas, es decir, harían la función de las restricciones; y la definición conceptual estaría encaminada a tratar de identificar información relevante para el dominio. En el siguiente ejemplo se muestra la estructura de una definición conceptual:

[**SUJETO**]
extrae: términos:
clases principales: clases modificadoras:
[**FORMA VERBAL**] verbo:
raíz:
modo:
[**COMPLEMENTO DIRECTO**]
extrae: términos:
clases principales: clases modificadoras:
[**COMPLEMENTOS INDIRECTOS**]
extrae: términos: preposiciones:
clases principales: clases modificadoras:
[**COMPLEMENTOS CIRCUNSTACIALES**]
extrae: términos: preposiciones:
clases principales: clases modificadoras:

El récord se dividió teniendo en cuenta las estructuras sintácticas y los campos se organizaron dentro de estas estructuras. Todo esto debido a la estrecha relación que mantiene la información lingüística con los constituyentes de la oración donde es hallada.

Los campos "clases principales" y "clases modificadoras" están destinados a almacenar las clases que representan a los conceptos semánticos, mientras que los campos "términos", "preposiciones", "verbo", "raíz" y "modo" toman valores que representan evidencias léxicas.

Es importante explicar la presencia del campo "extrae", que es un campo común en todos los constituyentes, excepto en el constituyente que representa la información de la forma verbal. Este campo es el que indica los conceptos semánticos propios del dominio que se deben de extraer dentro de cada constituyente sintáctico. La información de este campo se determinó teniendo en cuenta el criterio de verificar si las clases principales de cada constituyente mantenían una relación de subclasificación con las clases "**Control**", "**Cultivos**" y "**Plagas**", que son las clases que representan a los conceptos semánticos que se deben extraer para el dominio. Por ejemplo, si en un constituyente de la definición conceptual se encuentran las clases principales "**Cultivos_Permanentes**" y "**Ácaros**", entonces los conceptos semánticos que se extraen de éste son los "**Cultivos**" y las "**Plagas**".

El motivo por el cual no se incluyó un campo "extrae" dentro del constituyente que representa a la forma verbal es que los verbos no representan información semántica.

Las definiciones conceptuales no sólo se usan para representar a las reglas, éstas también se utilizan con el propósito de representar las evidencias semánticas y léxicas propias de cada una de las oraciones presentes en los textos que se usaron, tanto para el entrenamiento como para la prueba.

Un aspecto importante a resaltar es, teniendo en cuenta este uso, que la información lingüística de un constituyente determinado no se representa si este constituyente no está presente en la oración o si los términos que conforman a dicho constituyente no representan información semántica a tener en cuenta para el dominio.

Para el caso en que la definición conceptual sea usada como regla, si todas las restricciones de ésta son satisfechas por una oración, entonces se crea un récord cuyos campos tomarían los valores de las palabras o frases que representan contenido semántico de importancia para el dominio y que se pueden clasificar usando las clases presentes en los campos "extrae" de cada constituyente. En este caso se dice que la definición conceptual o la regla **cubre** la oración.

VII. Construcción de las reglas. Proceso de generalización

Las reglas para el análisis de textos de SEISAV son inducidas automáticamente a partir de un entrenamiento que involucra a textos que presentan oraciones previamente anotadas por un especialista en el dominio, es decir, a partir de instancias de entrenamiento. El objetivo de este entrenamiento es lograr reglas lo suficientemente generales que obtengan una precisión elevada en la extracción de información relevante para el dominio y que, a la vez, no estén tan cargadas de restricciones para lograr un funcionamiento correcto.

SEISAV comienza este proceso de inducción de reglas (o proceso de generalización) proponiendo una representación de cada instancia en el conjunto de entrenamiento, a partir de una definición conceptual que funcione como una regla. Claro, estas reglas serían demasiado específicas y sólo serían capaces de identificar información relevante en las instancias que las originaron, ya que solamente ellas serían capaces de cumplir las restricciones de cada regla a la que dieron lugar, es decir, cada regla solamente podría **cubrir** a la instancia que la originó. Por esto, para que estas reglas puedan ser usadas para extraer información en textos que no se utilizaron para el entrenamiento, deberían lograr un nivel de generalización mayor en sus restricciones que le permitiera a la regla cubrir una mayor cantidad de instancias.

Para lograr definiciones conceptuales más generales, SEISAV cuenta con un algoritmo de generalización que comienza seleccionando una definición conceptual inicial de las creadas a partir de

las instancias de entrenamiento y propone, para ésta, una definición más general que la que representa basándose en la definición conceptual más similar a la seleccionada. Para hallar esta definición conceptual más similar se consideró a todas las definiciones que tuvieran iguales valores en los correspondientes campos "extrae", o lo que es lo mismo, a todas las definiciones que extrajeran iguales conceptos semánticos, y se tuvo en cuenta la aplicación de una métrica de similitud.

La métrica de similitud usada se basa en la cantidad de relajaciones que se necesitan efectuar en las restricciones para lograr una generalización de ambas definiciones. La definición conceptual más similar a otra sería la que menor número de relajaciones necesitaría efectuar. Por lo tanto, una generalización para dos definiciones conceptuales se obtendría relajando lo más posible las restricciones de éstas para obtener una definición que cubra a las instancias representadas por ambas definiciones conceptuales.

VIII. Fase de prueba para las reglas inducidas

La fase de prueba consiste en aplicar las reglas a un conjunto de textos en los cuales los conceptos a extraer se encuentran previamente especificados, es decir, un conjunto de textos formados por oraciones previamente anotadas por un experto en el dominio, en las que todas poseen información relevante para el dominio. El objetivo de este proceso es medir el rendimiento de las reglas que se construyeron para ofrecer un estimado de la eficiencia con que ellas pueden operar.

Este proceso consiste en construir, para cada oración en el conjunto de prueba, una definición conceptual que represente las evidencias lingüísticas y especifique las extracciones que realiza cada una de ellas. Posteriormente, seleccionando una a una estas definiciones conceptuales, se busca cual es la regla que cubre con un menor valor de la métrica de similitud a dicha definición conceptual. Si la regla encontrada no realiza las mismas extracciones por constituyente que la definición conceptual, entonces se cuenta como que las reglas realizaron una extracción errónea. Si las extracciones coinciden, entonces se cuenta como que las reglas realizaron una extracción correcta. Si no se encuentra una regla que cubra la definición conceptual también se cuenta como que las reglas realizan una extracción errónea.

Después de aplicadas las reglas a cada una de las definiciones conceptuales se procede a calcular el rendimiento de estas reglas teniendo en cuenta el parámetro *recobrado.*
Se realizaron diferentes experimentos variando el tamaño del texto destinado al entrenamiento para observar como influía éste en el rendimiento de las reglas. Por esta razón los conjuntos de reglas obtenidas en cada experimento se sometieron de forma independiente a una fase de prueba. Los rendimientos obtenidos en estos experimentos se consideran bajos debido a los escasos materiales que se pudieron procesar para el entrenamiento del sistema.

IX. Comparación de SEISAV con CRYSTAL

Este trabajo se realizó usando como referencia al sistema CRYSTAL desarrollado en la Universidad de Massachussets. CRYSTAL fue un sistema implementado para inducir automáticamente un conjunto de reglas de análisis de textos para un dominio específico, a partir de ejemplos de entrenamiento. Este sistema construye reglas que se acercan en rendimiento a reglas construidas manualmente por un especialista en el dominio.

Uno de los aspectos más importantes en la construcción de un SEI es contar con un algoritmo que analice eficientemente un conjunto bastante grande de posibles reglas para poder obtener, mediante un proceso de generalización, reglas lo suficientemente expresivas que logren un alto rendimiento en el proceso de extracción de información de textos sujetos a escritura sin restricciones.

Primeramente hay que aclarar que CRYSTAL está dirigido a operar en textos escritos en lengua Inglesa, mientras que SEISAV se diseñó para trabajar con textos completamente en Español. Este es uno de los cambios más importantes, ya que los constituyentes que presentan las definiciones conceptuales son completamente diferentes en ambos sistemas. Por ejemplo, CRYSTAL usa constituyentes como *Subject*, *Verb*, *Object* y *Relative Object*, mientras que SEISAV trabaja con los constituyentes *Sujeto*, *Verbo*, *Complemento Directo*, *Complementos Indirectos* y *Complementos Circunstanciales.*

Otros cambios realizados se encuentran en los campos de los constituyentes. CRYSTAL utiliza una mayor cantidad de restricciones y, en el caso de la restricción perteneciente al modo del verbo, incluye dos posibles valores más que SEISAV.

Como se puede apreciar, SEISAV no distingue entre términos principales y términos modificadores, sí lo hace para las clases semánticas, pero no incluye la restricción de éstas en conjunto. En el caso de la restricción perteneciente al modo, solamente incluye los valores de afirmativo o negativo, no cuenta con los de activo o pasivo.

Tratar de utilizar las mismas especificaciones con que fue construido un SEI para un dominio en la confección de otro SEI en un dominio diferente puede resultar en la posterior extracción de información que no sea de interés para este nuevo dominio. Así mismo, como la tecnología moderna no está al alcance de lograr rendimientos semejantes a los de los humanos, tampoco está lo suficientemente desarrollada como para que se logre construir un sistema de propósito general que sea capaz de extraer eficientemente información en cada dominio a que se aplique.

Conclusiones

Este sistema de extracción de información para el contexto de control de plagas de plantas puede convertirse en una herramienta útil para el procesamiento de la información en un centro de información especializado en sanidad vegetal.

Para lograr su completa y correcta explotación, es necesario procesar una cantidad de textos mucho mayor que la analizada en el ensayo, para entrenar mejor al sistema y conseguir unos buenos parámetros de rendimiento. De este modo, el sistema contribuiría con la selección de los documentos relacionados con la temática así como con su indización.

Bibliografía

- **Appelt, D. E. y D. J. Israel**. 1999. «Introduction to Information Extraction Technology. A tutorial prepared for IJCAI-99». Artificial Intelligence Center. SRI International. http://www.ai.mit.edu/people/jimmylin/papers/intro-to-ie.pdf
- **Glickman, O. y Rosie Jones**. 1999. Examining Machine Learning for Adaptable End-to-End Information Extraction Systems. [ml4ie.glickman&jones.pdf] AAAI-99. Workshop on Machine Learning for Information Extraction. USA. http://www.aaai.org/ Conferences/National/1999

- **Hernández, Z., J. Pérez y O. Santana**. Extracción de Información._ España: Universidad de las Palmas de Gran Canaria. Grupo de Estructuras de Datos y Lingüística Computacional. http://www.gedlc.ulpgc.es/docencia/ seminarios/pln/Extraccion_de_informacion/sld032.htm
- **Morales, Eduardo**. Inteligencia Artificial. México: Instituto Tecnológico y de Estudios Superiores de Monterrey. Campus Morelos, 1999. w3.mor.itesm.mx/~emorales/Cursos/RdeC/node10.html.
- **Ruíz, Rafael**. El análisis documental: bases terminológicas, conceptualización y estructura operativa. España: Universidad de Granada, 1992.
- **Soderland, S.; D. Fisher, J. Aseltine y Wendy Lehnert**. 1995. Crystal: Inducing a Conceptual Dictionary. Proceedings of the fourteenth International Joint Conference on Artificial Intelligence. IJCAI-95. http://www-nlp.cs.umass.edu/ciir-pubs/soderland_ijcai95.pdf
- **Soderland, S**. 1997. Learning text analysis rules for domain-specific natural language processing. Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of Doctor of Philosophy. Computer Science Department at the University of Massachusetts.
- **Universidad de Sheffield**. Information Extraction. Grupo de Procesamiento del Lenguaje Natural. Departamento de Ciencias de la Computación. 2002. http://nlp.shef.ac.uk/research/areas/ie.html

SOBRE LA AUTORA:

Ingrid Paz García

Ciudad de la Habana, Cuba (1971). Ingeniero. Ha trabajado en el Centro de Información y Documentación del Instituto de Investigaciones de Sanidad Vegetal realizando actividades relacionadas con la especialidad, contribuyendo al perfeccionamiento de los procesos, productos y servicios de la biblioteca en el campo de la investigación científica. Actualmente trabaja en la Dirección Nacional del Grupo de Turismo Gaviota SA. en la Dirección de Informática y Comunicaciones, administrando la pagina web y la intranet, asi como evaluando sistemas de aplicaciones de gestión de información y del conocimiento.

c.e: ingrid@gaviota.gav.tur.cu