Sistema de Consultas en Lenguaje Natural para Bases de Datos

José Manuel Soto Corzo, David Díaz Portillo, José Antonio Cruz Zamora

Instituto Tecnológico de Apizaco
Av. Instituto Tecnológico de Apizaco s/n
alestad_raziel@hotmail.com, daviddp84@hotmail.com, tonocz2001@yahoo.com.mx

Abstract. Este trabajo se plantea el desarrollo de un módulo de consulta en lenguaje natural (en forma escrita) a una base de datos relacional empleada en el sistema ITASS.

Dicho módulo involucra los procesos propios de un sistema de consultas en lenguaje natural a bases de datos relacionales; análisis sintáctico, análisis semántico, análisis morfológico y la interpretación de la consulta a un lenguaje formal [Sopeña, 1994]. El propósito del proyecto mencionado es crear una herramienta que permita la recuperación de información de un sistema de bases de datos de tipo relacional por usuarios no especializados sobre el dominio propio del sistema.

Palabras Clave: Bases de datos, SQL, Procesamiento de Lenguaje Natural (PLN), Inteligencia Artificial (IA), Análisis Sintáctico, Semántico y Morfológico, Modelo Relacional.

1. Introducción

Desde la perspectiva de la inteligencia artificial (IA), el estudio del lenguaje natural tiene dos objetivos:

1. Facilitar la comunicación con la computadora para que accedan a ella usuarios no especializados.

2. Modelar los procesos cognoscitivos que entran en juego en la comprensión del lenguaje para diseñar sistemas que realicen tareas lingüísticas complejas (traducción, resúmenes de textos, recuperación de información, etc.)

Hay problemas en los que interesa fundamentalmente el primer objetivo. Lo que se desea es conseguir un intérprete para una clase de aplicaciones en un dominio restringido, que haga de traductor entre el ordenador y el usuario. El intérprete realiza dos tareas: una de reconocimiento de la instrucción del usuario, otra de generación de una expresión equivalente en un lenguaje formal que utilice la computadora para la aplicación. Este enfoque modela el lenguaje como una herramienta de comunicación

(usuario-sistema) sobre conjuntos de información tipificada y restringida, por tanto, solo tratara el subconjunto del lenguaje natural que describe los aspectos significativos en ese dominio. La comprensión de una consulta se plantea de este modo como un proceso de detección de los datos solicitados sobre el dominio de la aplicación.

El segundo objetivo plantea el lenguaje como objeto de estudio, y la comprensión como un proceso complejo en que intervienen grandes cantidades de conocimiento de naturaleza diferente (morfología, sintaxis, semántica, pragmática) y mecanismos de tratamiento variados (de comparación, búsqueda, inferencia aproximada, deducción, etc.). [Meya, 1986]

Este trabajo abarca lo referente al primer objetivo, implementando un módulo de consultas a un sistema de control que cuenta con una base de datos relacional.

En la sección 2, se menciona el proceso de análisis sintáctico de la instrucción en lenguaje natural; la sección 3, considera el proceso de análisis semántico de la instrucción en lenguaje natural; la sección 4, describe la metodología utilizada para el análisis morfológico de los elementos constituyentes clasificados por medio de dos autómatas de estados finitos; 5, describe la metodología utilizada para la interpretación de los elementos constituyentes a sus equivalentes a un lenguaje formal de consulta; y finalmente, se muestran las conclusiones.

2. Análisis sintáctico

De todos los niveles de análisis expuestos, la sintaxis ha sido durante mucho tiempo y aún sigue siendo el nivel al que la lingüística le ha prestado mayor atención. Está casi exclusiva atención se justifica por dos razones principales en cuanto al tratamiento automático del lenguaje natural:

1. El procesamiento semántico funciona sobre los constituyentes de la oración. Si no existe un paso de análisis sintáctico, el sistema semántico debe identificar sus propios constituyentes. Por otro lado, si se realiza un análisis sintáctico, se restringe enormemente el número de constituyentes a considerar por el semántico, mucho más complejo y menos fiable. El análisis sintáctico es mucho menos costoso computacionalmente hablando que el análisis semántico (que requiere inferencias importantes). Por tanto, la existencia de un análisis sintáctico conlleva un considerable ahorro de recursos y una disminución de la complejidad del sistema.

2. Aunque frecuentemente se puede extraer el significado de una oración sin usar hechos gramaticales, no siempre es posible hacerlo. [Rich & Knight, 1994].

La propuesta para el análisis sintáctico es el uso de un diccionario de palabras validas para el dominio de la aplicación. Dicho diccionario esta constituido por una tabla en una base de datos que contiene palabras propias del dominio. La estructura gramatical de la instrucción en lenguaje natural es descompuesta en palabras y símbolos de puntuación tomados cada uno de ellos como elementos constituyentes durante la fase de análisis semántico. Estos elementos son comparados con el diccionario, y en caso de ser validos se almacenan en forma de lista para servir de entrada al módulo de análisis semántico. De no encontrarse como valida alguna palabra, el usuario podrá añadir nuevos elementos al diccionario o redefinir su instrucción.

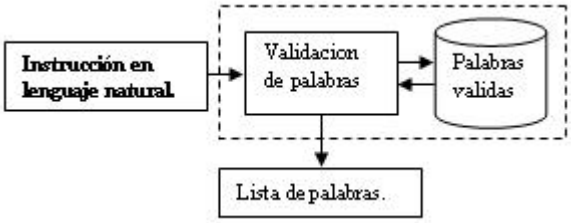

Fig. 1. Proceso de análisis sintáctico.

Se considera que, para obtener un análisis lingüístico válido se necesita un gran nivel de interacción entre los tradicionalmente separados módulos de análisis sintáctico y semántico.

Desde esta perspectiva, la función de un análisis de constituyentes se reduciría a generar la estructura base sobre la que podemos hacer funcionar otros módulos del sistema, aunque por supuesto, el analizador deberá recurrir a otros módulos para poder generar la estructura correcta. Lo más usual, sin embargo, es quedarse en el simple análisis de constituyentes. Por supuesto, el análisis no debe quedarse en las funciones semánticas básicas, en realidad, debería extenderse hasta el conocimiento pragmático y conocimiento del mundo o conocimiento de sentido común. Pero debido a la complejidad de este proceso el sistema propuesto no realiza un análisis pragmático.

3. Análisis Semántico

El análisis semántico se refiere a la detección del significado de cada elemento constituyente dentro de la oración. Para llevar a cabo este proceso se considera la división de una consulta en tres bloques:

1. Especificación de datos requeridos.
2. Definición del origen de dichos datos.

3. Especificación de condiciones de búsqueda.

Dichos bloques están definidos en el lenguaje formal de consulta (SQL) por las palabras: SELECT, FROM y WHERE respectivamente.

El módulo de análisis semántico toma como entrada la lista de palabras generadas por el módulo de análisis sintáctico y la somete a un proceso iterativo de revisión basado en inferencias para etiquetar los elementos constituyentes de acuerdo al bloque al cual corresponden, quedando las etiquetas de la siguiente manera:

Table 1. Etiquetas aplicadas a los elementos constituyentes durante el análisis semántico.

Bloque	Etiqueta
Datos solicitados	1
Origen de los datos solicitados	2
Condiciones	3

Una vez etiquetada la lista de elementos constituyentes, esta es dada como entrada a un submódulo de Separación por bloques, el cual en base a la etiqueta de cada elemento forma una cadena de texto para cada uno de los bloques mostrados en la tabla 1. Estas cadenas de texto sirven de entrada a un modulo de reestructuración, que reordena los elementos constituyentes de los bloques a fin de poder ser interpretados correctamente.

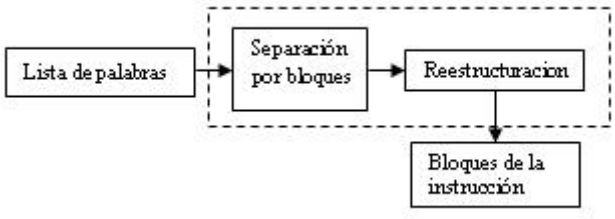

Fig. 2. Proceso de análisis semántico.

Si durante alguna fase del análisis semántico se detecta un error en la instrucción, el sistema envía un mensaje de error y permite al usuario redefinir su consulta.

4. Análisis Morfológico

El análisis morfológico se refiere a la revisión de la correcta estructura de la instrucción. Para ello se han definido los bloques requeridos en una consulta dentro de la instrucción en lenguaje natural, y se han separado en cadenas de texto para servir de entrada a este modulo. El análisis morfológico de los bloques se hará por medio de dos autómatas de estados finitos, uno para el bloque de datos solicitados y otro para el bloque de condiciones, así como un proceso de concatenación de los

elementos contenidos en las cadenas de texto que contienen los elementos constituyentes de cada bloque.

Table 2. Autómata de estados finitos para el bloque de datos solicitados.

Estado 1	Estado 2
Campo de la base de datos.	Indicador de conjunción.

Donde el indicador de conjunción podrá ser una coma o la palabra “y”.

Table 3. Autómata de estados finitos para el bloque de condiciones.

Estado 1	Estado 2	Estado 3	Estado 4
Campo de la base de datos.	Operador de comparación.	Valor de restricción.	Indicador de conjunción.

Donde el operador de comparación podrá ser toda palabra que corresponda con alguno de los operadores de comparación utilizados en SQL (<, >, >=, <=, like); el valor de restricción podrá ser toda palabra que denote un valor restrictivo para el campo de la base de datos, pudiendo ser este un numero, una cadena de texto o una fecha; el indicador de conjunción podrá ser una coma o la palabra “y”.

4.1 Autómata de estados finites

La representación de reglas morfológicas mediante un autómata de estados finitos tiene la ventaja de su fácil implementación y un procesamiento altamente rápido por lo que se le debe considerar como un método optimo para el procesamiento morfológico de cualquier índole.

La evaluación de la instrucción se hace mediante un autómata de estados finitos. El proceso va agrupando elementos hasta que encuentra en el diccionario de morfemas una entrada igual. En este punto el autómata ha reconocido un posible bloque. Si luego los rasgos que tiene en el diccionario tales elementos, le reconocen como adecuado para el tipo de estado para el cual se encuentra el autómata se consume tal estado y se pasa al siguiente.El análisis morfológico arroja como salida un cadena de texto que servirá de entrada al modulo de interpretación de la instrucción al lenguaje formal de consulta SQL.

Si durante alguna fase del análisis morfológico se detecta un error en la instrucción, el sistema envía un mensaje de error y permite al usuario redefinir su consulta.

Modulo de Interpretación

Una vez evaluado y confirmado la estructura sintáctica y semántica de la instrucción, deberá obtenerse un equivalente en un lenguaje formal de consulta que permita la obtención de la información requerida. Para la obtención de la instrucción final en SQL se propone el uso de un almacén de datos que contenga los comandos equivalentes de SQL a los constituyentes que conforman la instrucción en lenguaje natural. El proceso consistirá en una búsqueda secuencial indexada de equivalencias dentro de una tabla que hará la función de un traductor. Dicha tabla de equivalencias contiene un listado de palabras y su equivalente en el lenguaje formal de consulta de acuerdo a la nomenclatura del constituyente evaluado.

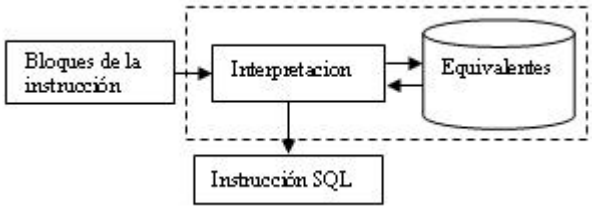

Fig. 3. Proceso de interpretación.

Conclusiones

El desarrollo de sistemas de procesamiento automático del lenguaje natural permite crear herramientas útiles para hacer más flexible la comunicación con la computadora por usuarios no especializados para la especificación de tareas, en este caso, para la recuperación de información. Así también permiten avanzar en el logro de una interacción hombre – maquina en forma natural.

Los sistemas de procesamiento del lenguaje natural conllevan un amplio proceso de análisis para su posterior interpretación a un lenguaje formal de computadora y su aplicación esta restringida a un dominio reducido, dependiendo este de la aplicación de que se trate.

El uso de esta metodología podría minimizar los procesos computacionales requeridos para la interpretación de una instrucción en lenguaje natural, pudiéndose desarrollar algoritmos generales, aplicables a diversos dominios.

Referencias

1. L. de Sopeña. ULS un sistema para interrogar a bases de datos relacionales. Centro científico IBM, Madrid, España. [Revista SEPLN]. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/1/1-articulo-8.pdf
2. Montserrat Meya. Análisis morfológico automático del español. Centro de investigación SIEMENS, Munich, Alemania. [Revista SEPLN]. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/1/1-articulo-4.pdf.
3. Rich, E., Knight, K. (1991). *Inteligencia artificial*, 2da edición. McGraw-Colina, Nueva York, Nueva York.
4. Grishman, R. (1986) *Computational Linguistics*. Cambridge: Cambridge University Press.
5. Julia Díaz García, Pilar Rodríguez Marín. PROGENES: La interfaz en Lenguaje Natural. [Universidad de Valencia, España]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/11/11-Pag41.pdf
6. Juan Barreras. Resolución de Elipses y Técnicas de Parsing en una Interficie de Lenguaje Natural. [Departamento de I+D NLU/ISS SA, España]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/13/13-Pag247.pdf
7. Julia Díaz García, Julio González Arroyo. El Formalismo Semántica en la Interfaz de Lenguaje Natural Progenes. [Universidad de Valencia, España]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/14/14-Pag119.pdf
8. Fernando Sánchez León. Desarrollo de un Etiquetador Morfosintactico para el Español. [Universidad de Madrid, España]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/17/17-Pag14.pdf
9. José F. Quesada. Un Modelo Robusto y Eficiente para el Análisis Sintáctico de Lenguajes Naturales Mediante Árboles Múltiples Virtuales. [Centro Informático de Andalucía, España]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/19/19-Pag14.pdf
10. Paloma Martínez, Ana García Serrano. Una Propuesta de Estructuración del Conocimiento para la Adquisición de Esquemas Conceptuales de Bases de Datos a partir de Textos [Universidad Carlos III y Politécnica de Madrid, España]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/21/21-Pag91.pdf
11. T. L. Soto, J. F, Quesada. Parsigns Strategies for a Spoken Language Processign System. [Universidad de Sevilla, España]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/23/23-Pag8.pdf
12. Javier Couto, Gustavo Crispino. Estructuración de Índices Gramaticales y Léxicos para la Extracción y Recuperación de Información. [Universidad de la Republica de Uruguay, Uruguay]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/25/25-Pag43.pdf
13. Nuria Gala Pavia .Using the Incremental Finite State Architecture to Create a Spanish Shallow Parser. [Xerox Research Centre Europe, France]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/25/25-Pag75.pdf

14. Pablo Gamillo Otero, Marie Laure Reinberger. Modelizacion de la Combinación Dinámica de Estructuras Lexicas [Loboratoire de Researche sur le Langage, Maison de la Researche, Francia]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/25/25-Pag83.pdf
15. Alicia Garrido, Amaia Iturraspe. A Compiler for Morphological Analyser and Generators Based on Finite State Transducters. [Universidad de Alcántara, España]. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/25/25-Pag93.pdf
16. Montserrat Marimon, Axel Theofilidis. Linguistic Processing Modules in ALEP for Natural Language Interfaces. Revista SEPLN. [Citado Septiembre 10, 2006]. Disponible en la World Wide Web: http://www.sepln.org/revistaSEPLN/revista/25/25-Pag129.pdf