ISSN: 1946-5351 ONLINE, 1948-2353 CD ROM Vol. 5 No. 3

SNL2SQL: Conversión de consultas en SQL al idioma Español

Dr. Ismael Esquivel Gámez[1], MC Rafael Córdoba Del Valle[2], LSCA Daniel González Espinoza[3], LSCA Eliana Ogarita Guadalupe López Collins[4]

Resumen.**Las interfaces de lenguaje natural a bases de datos (ILNBDs) permiten a los usuarios finales formular sus necesidades de información en lenguaje natural y, sin la intervención de personal del área de informática, cubrir sus necesidades de información. Uno de los mecanismos para medir la efectividad de la interfaz es el grado de proximidad entre la consulta original del usuario y la proporcionada por un medio de retroalimentación automatizado. Este informe presenta un módulo de conversión de consultas en SQL al idioma español, formando parte de una interfaz de lenguaje natural a base de datos, aun en construcción, denominada SNL2SQL.**
Palabras clave: **SQL, ILNBD, retroalimentación**

Introducción

El proyecto que se presenta, conjuga las áreas de lingüística computacional y de bases de datos. De la primera se usa la traducción automática, la cual investiga el uso de software para traducir texto o habla de un lenguaje natural a otro. De la segunda, se aplica la explotación de datos mediante el lenguaje estructurado de consultas (SQL). La solución que se plantea corresponde a lo que Androutsopoulos y otros (1995) refieren como interfaces de lenguaje natural a bases de datos (ILNBD). Aunque ya existen desarrollos en nuestro idioma, todavía no han cumplido con las expectativas originalmente planteadas.

Esquivel (2011) indica que: El SQL como estándar para el manejo de bases de datos relacionales, ha impactado en múltiples área del mercado del cómputo. El producto DB2 de IBM domina el manejo de datos en equipos de gran tamaño. Oracle con su producto, domina el área de sistemas para computadoras basadas en Unix. SQL server de Microsoft predomina en las soluciones orientadas a grupos de trabajo y departamentos, que usan el sistema operativo Windows y MySQL domina el mercado de bases de datos de código abierto. SQL se ha aceptado como una tecnología apropiada para el procesamiento de transacciones en línea, además de que aplicaciones de minería de datos y de almacenes de datos basadas en SQL son la base para el descubrimiento de patrones de consumo y por tanto, el ofrecimiento de mejores productos y servicios. En Internet, las bases de datos manejadas con SQL son la piedra angular para ofrecer servicios y productos más personalizados, que son la clave de crecimiento del comercio electrónico.

SQL se ha convertido en el lenguaje de base de datos universal 'de facto' y aun cuando cada Sistema Gestor de Bases de datos (SGBD) le ha agregado sus particularidades, generando un sinnúmero de dialectos, en el presente proyecto se usa el SQL estándar ANSI-92, que es soportado por la mayoría de los motores de bases de datos relacionales.

[1] Doctor en Tecnología de Información y Análisis de Decisiones. iesquivel@uv.mx
[2] Maestro en Ciencias de la Computación. rcordoba@uv.mx
[3] Licenciado en Sistemas Computacionales Administrativos. danny.saga@hotmail.com
[4] Licenciada en Sistemas Computacionales Administrativos. ecollins.t16@hotmail.com

ISSN: 1946-5351 ONLINE, 1948-2353 CD ROM Vol. 5 No. 3

Proyecto SNL2SQL

La interfaz de lenguaje natural español a SQL (SNL2SQL) es una iniciativa subvencionada por el gobierno federal mexicano a través de la SEP/PROMEP, que tiene por objeto la generación automática de informes, mediante un sistema de traducción de peticiones hechas en lenguaje natural a comandos en SQL. Hasta el momento, el desarrollo del sistema ha logrado resultados prometedores (Esquivel et al., 2010), que corresponden a:

- Manejo de expresiones temporales (Fechas)
- Traducción de términos de agrupamiento
- Tratamiento de expresiones estadísticas

Sin embargo, dada la complejidad del proyecto se requiere de tiempo adicional para trabajar con los módulos: De filtrado de grupos, de manejo de sub-preguntas y procesamiento de expresiones horarias. Adicionalmente, y de la interacción con personal de informática de las empresas vinculadas, se ha encontrado que para garantizar los mejores resultados, el sistema necesita traducir el comando SQL obtenido a una pregunta en lenguaje natural, antes de proceder a obtener el informe, a fin de verificar si el sistema "ha entendido" correctamente la petición. También se ha detectado que los usuarios acostumbran realizar peticiones con una gran economía de palabras, pues una vez que emiten la primera pregunta, requieren que el sistema entienda las subsecuentes sin tantas palabras (preguntas elípticas).
Estas dos últimas funcionalidades permitirán una mejor interacción con el sistema y una mejoría en la efectividad lograda hasta el momento. Por lo tanto, para una segunda fase del proyecto se ha planteado la necesidad de desarrollar y probar los módulos que aparecen sin sombrear en la figura 1.

Figura 1. Módulos del SNL2SQL

En este documento se reportan las tareas del desarrollo y prueba del módulo de retroalimentación al usuario final sobre su petición original a la ILNBD, como un medio para medir la efectividad del sistema en lo general y lo específico. Adicionalmente, el módulo puede servir para fines didácticos, al aplicarse durante la enseñanza del SQL.

Diccionarios

El módulo utiliza dos diccionarios, el principal llamado también de Dominio y el secundario, denominado de Contenido. El primero se construyó a partir de un diccionario de sinónimos y del metadata de la base de datos, apoyado en el trabajo de Pazos et. al (2005). Del metadata se extraen por cada tabla: nombre, descripción de la misma y los datos relativos a sus columnas (nombre, tipo de dato, tamaño, permisibilidad de valores nulos y descripción de la misma). A partir del procesamiento de ésta última, se

ISSN: 1946-5351 ONLINE, 1948-2353 CD ROM Vol. 5 No. 3

obtienen los sustantivos asociados con las columnas y tablas que en su descripción, los contengan a ellos o sus sinónimos. Para este trabajo, el diccionario de sinónimos se ha construido manualmente usando el diccionario en línea que se provee en Rodríguez y Carretero (2008). Algunas entradas del citado diccionario se muestran en la figura 2.

Tabla	Columna	Calificador	Sinónimos	Formas Verbales
Empl	Birthdate	Nacimiento Nació	Alumbramiento, origen, principio	Nació, nacieron,
Empl	Hiredate	Ingreso Ingresó	Entrada, contratación, alta, admisión	Contratados, ingresaron, iniciaron, contrataron,

Figura 2. Muestra del diccionario de dominio

Para atributos que manejan un conjunto de valores válidos (dominios), principalmente para aquellos que manejan valores breves (Ej. Género: 'F', 'M'), es necesario capturar los sinónimos de dichos valores (Ej. 'MUJERES','FÉMINAS','FEMENINO','HOMBRES','VARONES','MASCULINO', etc.), en un archivo denominado Diccionario de Contenido. Éste sirve para facilitar el procesamiento, principalmente de la clausula WHERE y con ello, se anima al uso de términos, propios del ámbito de la empresa, donde se aplique el presente módulo.

Adicionalmente, para las funciones de acumulado del SQL y para las cláusulas de agrupamiento y ordenamiento, se han creado diccionarios de expresiones homólogas para enriquecer el contenido de las oraciones producidas. El módulo en cada conversión, elige aleatoriamente una expresión distinta del correspondiente diccionario.

Algoritmo principal

Para realizar una consulta en SQL, se utiliza el comando SELECT cuya sintaxis aparece en la tabla 1 y en la cual, las cláusulas entre corchetes son opcionales.

Cláusula	Componentes	Significado
SELECT	Funciones, Columnas, Expresiones, Letreros	Muestra
FROM	Tabla(s) de la Base de datos	De la(s) tabla(s)
[WHERE]	Ecuaciones de vinculación, Columnas ó expresiones, operadores y valores	Donde
[GROUP BY]	Columnas de agrupamiento	Agrupados por
[HAVING]	Funciones, columnas, operadores y valores	Como condición
[UNION]		Unida a
[ORDER BY]	Columnas de ordenamiento	Ordenados por

Tabla 1. Sintaxis del comando SELECT

Como se aprecia en la tabla 1, en todo comando SELECT aparecen términos constantes y variables. Los constantes corresponden a cláusulas y sub-cláusulas del comando, mientras que las variables, a tablas, columnas y datos. Algunas expresiones constantes se traducen conforme al diccionario de expresiones homólogas citado anteriormente. Las variables se convierten mediante el diccionario de dominio y los datos, si necesario, usando el diccionario de contenido.

De acuerdo a la clasificación de consultas hechas por González y otros (2007), consistente de 6 tipos y tomando como base, una serie de comandos SELECT previamente verificados en su sintaxis, se procesaron

ISSN: 1946-5351 ONLINE, 1948-2353 CD ROM Vol. 5 No. 3

para obtener las oraciones en español que permitieran verificar la cercanía de dichas oraciones a las peticiones originales. Para ello, se ha formulado el procedimiento siguiente:

1. Se determinan las cláusulas y sub-cláusulas presentes para su conversión mediante un arreglo de términos base.
2. Luego se detectan las variables para su búsqueda en el diccionario de dominio y posterior traducción, usando para ello, el término calificador o alguno de sus sinónimos.
3. Enseguida, al encontrar los términos correspondientes a datos, comúnmente asociados a la cláusula WHERE o HAVING, se verifica que existan en el diccionario de contenido para su eventual conversión.
4. Una vez convertidos todos los componentes se procede a integrar la oración completa para su presentación.

Ejemplo:

```
SELECT MAX(SALARY)
WHERE JOB = 'DESIGNER'
```

Se traduce a:

Muestra máximo de salario donde puesto sea igual a diseñador

Implementación del Módulo

Se utilizó como base de datos de prueba, la mostrada en la figura 3. Por otro lado, el lenguaje de programación utilizado para el desarrollo fue REXX (REstructured eXtended eXecutor), un lenguaje de programación desarrollado en IBM por Michael Cowlishaw del que existen numerosas implementaciones disponibles con código abierto, conforme a Mertz (2004). Se trabajó con la versión 3.4 del intérprete Regina Rexx para Windows y se eligió porque cuenta con un gran conjunto de funciones, especialmente de tratamiento de textos y de fechas. Otra ventaja crucial de usar REXX es su orientación a multiplataforma, ya que con pequeñas adaptaciones puede ejecutarse en cualquier computadora y sistema operativo.

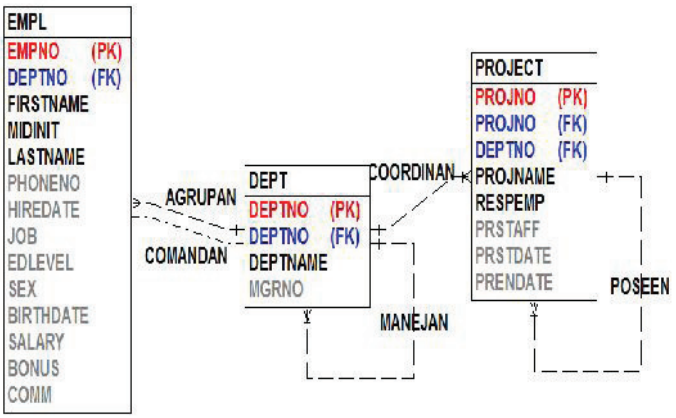

Figura 3. Diagrama lógico relacional de la BD de prueba.

ISSN: 1946-5351 ONLINE, 1948-2353 CD ROM Vol. 5 No. 3

Resultados

Se escogieron 30 consultas a la Base de datos que representaban la mayoría de las realizadas por usuarios finales. Para su prueba reiterativa, se guardaron en archivos textuales designados por queRynn.txt, donde nn representaba el número de consulta a procesar. Para probar el módulo, se pedía el número de consulta y enseguida se desplegaba la oración traducida, como se aprecian varios casos en la figura 4.

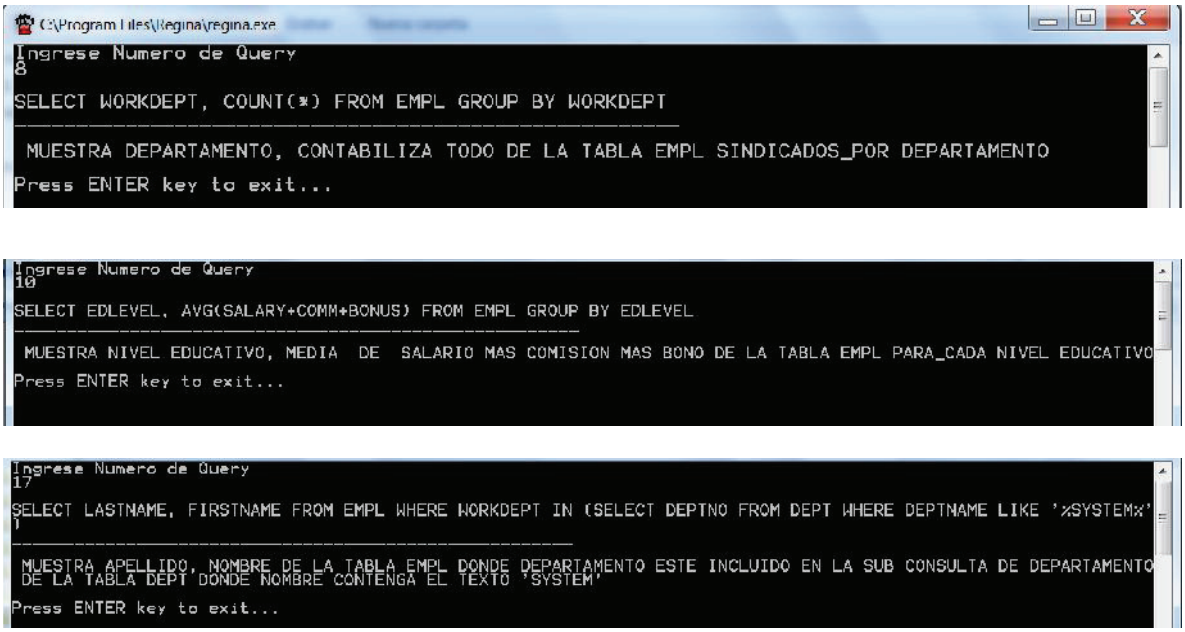

Figura 4. Muestra de la salida del proceso de conversión

Entre los tipos de consulta más complejos en su conversión se encontraban las que incluían sub-consultas (SUBQUERY), pero sobre todo aquellas que eran del tipo correlacional. Afortunadamente, se ha logrado superar la meta inicialmente fijada, de conseguir un 80% de efectividad en la traducción.

Sin embargo, hasta el momento se realiza la conversión directa sin realizar al final, el ajuste necesario para entregar una oración más parecida a la petición emitida por una persona.

De modo que un resultado como:
MUESTRA APELLIDO, SALARIO DE LA TABLA EMPL DONDE DEPARTAMENTO SEA EQUIVALENTE A CUALQUIER VALOR DE LA SUBCONSULTA DE DEPARTAMENTO DE LA TABLA DEPT DONDE NOMBRE ESTE INCLUIDO EN PLANEACION, SERVICIOS DE SOPORTE, OPERACIONES

Cambie por:
Despliegue el apellido y salario de quienes trabajan en los departamentos de Planeación, Servicios de soporte y Operaciones.

También, se ha presentado el módulo a dos grupos de alumnos de licenciatura en Sistemas Computacionales, para apoyarles en el aprendizaje del comando SELECT de SQL, de modo que luego de revisar el comando a traducir y responder al cuestionamiento sobre la información que provee el comando, ejecutan el módulo para comparar el resultado con su propia respuesta.

ISSN: 1946-5351 ONLINE, 1948-2353 CD ROM Vol. 5 No. 3

Conclusiones

Los resultados que se han conseguido hasta ahora, animan al desarrollo de la siguiente versión que permita una mayor fidelidad en la traducción con respecto a peticiones hechas por usuarios. Así mismo, promete que al término del desarrollo de la interfaz SNL2SQL, los beneficios a conseguir serán de gran impacto al reducirse el tiempo de generación de informes y la autonomía proporcionada a los usuarios finales de cualquier empresa o institución, que cuente con sistemas de información fundamentados en bases de datos relacionales.

Referencias

Androutsopoulos L., et al. (1995). *Natural Language Interfaces to Databases - An Introduction*, Journal of Natural Language Engineering, Vol. 1, 1995, pp. 29-81.

Esquivel, Ismael (2011). *SQL interactivo*. Editorial Académica Española. ISBN 978-3-8465-6468-4

Esquivel, I., et al. (2010). *Conversión de expresiones temporales de Español a SQL*. Revista de la Sociedad Española de Lenguaje Natural, Núm. 45, septiembre 2010, pp 229-237. ISSN 1135-5948.

Esquivel, I., et al. (2010). *Translation of Spanish Statistics Expressions to SQL*. Advances in Soft Computing Algorithms. Research in Computing Science 49, 2010, pp. 39-46. ISSN: 1870-4069.

González J., Pazos, R. A., Gelbukh, A., Sidorov, G., Fraire, H., Cruz, C. (2007). *Prepositions and conjunctions in a natural language interfaces to databases.* In: Thulasiraman, P., et al. (eds.) ISPA 2007 Workshops. LNCS Vol. 4723, pp. 173-182, Springer, Heidelberg

Mertz David (2004). *Rexx for everyone: Scripting with Free Software Rexx implementations.* http://www.ibm.com/developerworks/library/l-rexx.html

Pazos R., Pérez J., González J. J., Gelbukh A., Sidorov G. y Rodríguez M. (2005). *A Domain Independent Natural Language Interface to Databases Capable of Processing Complex Queries.* MICAI 2005: Advances in Artificial Intelligence. MICAI 2005: 833-842

Rodríguez, S. y Carretero, J. (2008). *COES: Herramientas para Procesamiento de Lenguaje Natural en Español.* http://www.datsi.fi.upm.es/~coes/interactivo/sinonimos.cgi