PROTEUS: UN SISTEMA MULTILINGÜE DE EXTRACCIÓN DE INFORMACIÓN

Antonio Moreno Sandoval
Cristina Olmeda Moreno
Ralph Grishman
Catherine Macleod
John Sterling

Department of Computer Science
Courant Institute of Mathematical Sciences
New York University
{morenoa,morenoc,grishman,sterling,macleod}@cs.nyu.edu

Resumen

El sistema PROTEUS (PROtotype TExt Understanding System) tiene como objetivo analizar e interpretar textos reales y mostrar un resumen de la información contenida en ellos de una forma estructurada (concretamente en forma de registro de base de datos). El sistema utiliza gramáticas del inglés y español de gran cobertura sintáctica, aunque la interpretación de las estructuras analizadas sólo se puede hacer dentro de dominios lingüísticos restringidos.

En esta comunicación presentaremos una aplicación práctica de PROTEUS: resumen de documentos periodísticos de información (no de opinión). Nuestro ejemplo está tomado de partes de agencias. Como es habitual en las bases de datos, la estructura ha sido predefinida en función de las necesidades de los usuarios. También se compararán los resultados obtenidos automáticamente por ambas gramáticas con los generados por un especialista.

1 ¿Qué es un sistema de extracción de información?

Grishman(1991) resume las características de este tipo de sistemas de la siguiente manera: son sistemas que convierten la información almacenada en forma textual (libros, revistas, informes técnicos...) en otra forma mucho más estructurada (por ejemplo, en registros de bases de datos). De esta manera se facilita grandemente el acceso a vastas colecciones de datos que sólo están accesibles en fondos documentales.

Es importante resaltar la diferencia entre extracción de información y recuperación de información. En el segundo caso, el objetivo del sistema es, en respuesta a una pregunta formulada por el usuario, buscar dentro del corpus y mostrar el/los documento/s relevante/s. Dentro de este tipo de sistemas documentales se pueden incluir tanto los que basan su búsqueda en la localización de palabras-clave dentro del texto (bases de datos documentales como GNOSIS), y aquellos que realizan algún tipo de procesamiento lingüístico parcial del texto (por ejemplo, el sistema FERRET desarrollado en Carnegie Mellon (Mauldin 1991)).

Los sistemas de extracción de información, en cambio, idealmente deberían analizar e interpretar el texto pues su cometido es «resumir» el contenido.

Debido a las características intrínsecas de cada sistema, los que están basados en palabras-clave son más rápidos pero menos precisos que los que procesan lingüísticamente el texto. Mauldin(1991) expone las razones de este hecho. Básicamente, los sistemas que buscan palabras se encuentran con los problemas de la polisemia y la sinonimia. Si uno busca documentos con palabras de múltiples significados se puede encontrar con documentos donde la palabra aparezca con un uso que no es el

consultado. En este caso se dice que se reduce la precisión (una de las dos medidas que se utilizan para medir la actuación (performance) de un sistema de recuperación). Es decir, se han recuperado más documentos de los que son relevantes.

La sinonimia reduce el recall, o proporción de documentos relevantes que se han recuperado. Si tenemos varias palabras o compuestos que describen el mismo concepto, cuando utilizamos una palabra concreta no podremos recuperar los documentos relevantes que no contengan esa palabra sino un sinónimo[1]

Por tanto, cualquier sistema documental que «comprenda» textos tendrá, en teoría, mejor actuación (es decir, sabrá reconocer con mayor precisión toda la información que se busca).

Para evaluar sistemas de extracción podemos utilizar técnicas muy parecidas. En primer lugar seleccionamos una muestra representativa del corpus y generamos manualmente el conjunto de «plantillas»[2] correctas, que compararemos con las plantillas generadas automáticamente por el sistema. Si contamos el número de campos en el conjunto de plantillas correctas, el número de campos en el conjunto de plantillas generadas por el sistema y el número de campos rellenados correctamente por el sistema obtendremos dos medidas de evaluación análogas a las usadas con los sistema de recuperación (Sundheim 1991):

$$\text{recall} = \frac{\text{núm. de campos rellenados correctamente por el sistema}}{\text{núm. de campos en las plantillas correctas}}$$

$$\text{precisión} = \frac{\text{núm. de campos rellenados correctamente por el sistema}}{\text{núm. de campos en las plantillas correctas}}$$

Hay que tener en cuenta a la hora de evaluar un sistema de extracción el tipo de subdominio lingüístico de la aplicación. Por ejemplo, se han utilizado algunos sistemas para crear bases de datos a partir de informes médicos. En estos casos el grado de exactitud tiene que ser muy alto. Por el contrario, en la mayoría de las aplicaciones es aceptable una exactitud menor.

Los sistemas prácticos actuales obtienen un recall de alrededor del 50% y una precisión algo mayor. Aún con estas limitaciones, si el volumen de texto es tan grande que no puede ser procesado por personas (o es demasiado costoso) extraer parte de la información es preferible a no conseguir nada.

2 Características de PROTEUS

La presente aplicación del sistema tiene tres compontentes principales: el analizador sintáctico, el analizador semántico y el generador de plantillas. La implementación se ha realizado en Allegro Common LISP y en la actualidad funciona en estaciones de trabajo SUN Sparc 1 y 2.

El analizador sintáctico es una evolución del Linguistic String Parser (Sager 1981) e incluye el parser, el analizador léxico y el compilador del Restriction Language. Es el elemento común de todas las aplicaciones del sistema y ocupa aproximadamente unas 4500 líneas de código. El analizador semántico fue desarrollado originariamente para otra aplicación y ha sido revisado; tiene 3000 líneas de código. El generador de plantillas ha sido creado especialmente para la aplicación actual y ocupa 1200 líneas de Common LISP.

---

[1] Se han dado soluciones parciales a ambos problemas. Los detalles y referencias se pueden encontrar en el mencionado artículo de Mauldin.

[2] En nuestro sistema denominamos «plantillas» a cada uno de los registros de la base de datos. Dichas plantillas tienen una estructura definida previamente (al igual que cualquier registro de una base de datos) con un número determinado de campos (*slots* en nuestra terminología) que hay que rellenar.

2.1 Estructura del sistema

El sistema PROTEUS realiza un análisis sintáctico completo de cada oración del texto, proporcionando una estructura regularizada para cada análisis. La estructura regularizada es la entrada para el análisis semántico, que reduce el mensaje a una estructura de papeles temáticos. Por último, el generador de plantillas traslada la interpretación a los campos de la base de datos.

Cada texto atraviesa cinco etapas en el proceso: análisis léxico, análisis sintáctico, análisis semántico, resolución de anáforas y generación de plantillas. Este proceso es básicamente secuencial ya que cada oración del texto pasa por las cuatro primeras etapas. Una vez que se ha interpretado el mensaje completo, se generan una o varias plantillas de acuerdo con la información que se haya extraido (Grishman et al., 1991). La figura 1 contiene una descripción general de los distintos componentes del sistema y la tarea que ejecuta cada uno de ellos.

A continuación, vamos a describir brevemente los distintos componentes del sistema.

El diccionario: contiene únicamente información morfosintáctica. En el sistema para el inglés, el diccionario de clases abiertas ha sido generado automáticamente a partir de la versión en SGML del *Oxford Advanced Learner's Dictionary («OALD»)* y tiene unas 10.000 entradas. Para el español, el diccionario es en la actualidad pequeño (unas 1.000 entradas) y está enteramente codificado a mano. Se está intentado conseguir algún diccionario electrónico del español para generar el lexicón automáticamente a partir de él y ampliar de esta forma el vocabulario.

(GRAPHIC IMAGE: NA)

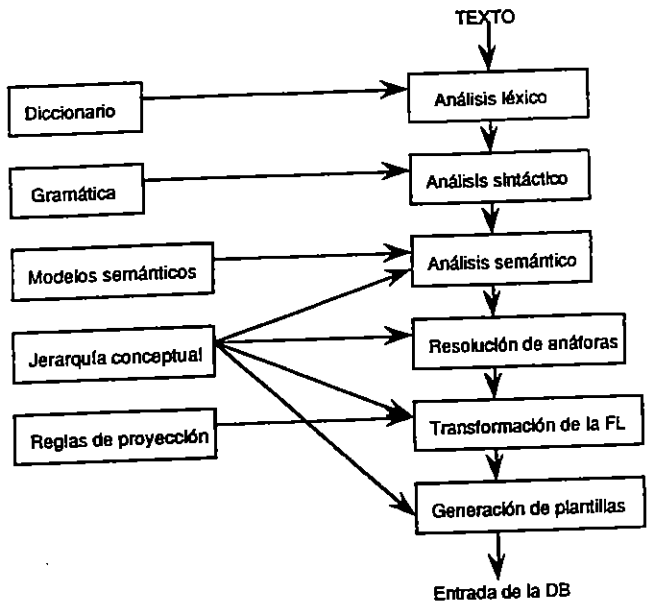

Figura 1. Estructura del sistema PROTEUS

El tratamiento morfológico es de tipo paradigmático. Para los casos regulares, el diccionario contiene una forma básica que se expande mediante macros hasta completar todas las formas del paradigma. Para los casos irregulares, el diccionario registra todas las formas que presentan una irregularidad; las restantes se generan de forma automática siguiendo los modelos regulares. El diccionario también incluye formas lexicalizadas y compuestos.

Las entradas típicas de diccionario para verbos, nombres y adjetivos tienen la forma:

```
(VERB :ROOT «DECLARAR»
      :ATTRIBUTES (REFLEXIVE)
      :OBJLIST (QUES INDOBJ-QUES PREDICATIVO))
```

(NOUN :FEM-SING «DECLARACIÓN»)
(ADJECTIVE :ROOT «AMERICANO»)

La gramática: el formalismo está basado en Linguistic String Grammar (Sager 1981). Es una gramática de estructura sintagmática independiente del contexto ampliada con restricciones escritas en un lenguaje procedural denominado Restriction Language. Admite el uso de rasgos y tiene una rutina para comprobar y realizar unificación de valores. El Lenguaje de Restricciones permite establecer con mucha exactitud la aplicación de las reglas.

La gramática cubre un gran subconjunto del español: coordinación, subordinación, elisión, clíticos, distintos órdenes de constituyentes. Está pensada para analizar textos periodísticos con oraciones de una longitud media de 35 palabras.

La jerarquía conceptual y los modelos semánticos: cada palabra está relacionada con un concepto mediante la relación *is-a* (es un ejemplo de...) y los distintos conceptos están también organizados en una estructura jerárquica. Las jerarquías no son completas ni universales sino que se crean específicamente para cada aplicación. Nuestro dominio actual son los informes sobre terrorismo[3]. Por ejemplo, la palabra «bomba» es un ejemplo del concepto |OBJETO EXPLOSIVO| (que contiene también otras palabras como «coche-bomba» o «carga explosiva») que a su vez es un tipo de |ARMA|, de tal forma que las restricciones seleccionales que se apliquen a |ARMA| se extienden por herencia a todas sus subclases.

Cuando una de las estructuras regularizadas producidas por el analizador sintáctico entra en el componente semántico, a cada palabra se le asigna su correspondiente clase semántica. A continuación, se intenta casar la estructura de la oración con alguno de los modelos semánticos. La misión de estos modelos es proyectar la estructura sintáctica en una estructura de papeles y de relaciones. Tomemos como ejemplo la oración: *El atentado de ETA se saldó con la muerte de dos transeúntes.* Existen tres modelos relevantes para la interpretación de esta oración. El primero de ellos representa el verbo «saldarse», que especifica

1) que el sujeto de este verbo debe pertenecer a la clase *suceso* y el papel que se le asigna en la estructura semántica es el de ***causa***,

2) que el objeto preposicional (y en concreto la preposición debe ser «con») tiene que pertenecer también a la clase de los sucesos y su papel es el de ***efecto.***

Cada uno de estos dos sucesos debe coincidir a su vez con alguno de los modelos definidos. En efecto, existe un modelo para «ataque» (cuyo sinónimo es «atentado») donde se establece que puede aparecer un modificador con la preposición «de» que pertenece a la clase *entidad-de-acción* y cuyo papel será el de ***agente*** del ataque. Por último, existe también un modelo para «muerte» en el que la preposición «de» introduce una *entidad-objetivo* que asume el papel de ***paciente*** de la acción.

En resumen, tenemos dos sucesos (una causa =»el atentado de ETA» y un efecto = «la muerte de dos transeúntes») que a su vez tienen una estructura interna (agente de la causa = «ETA»; paciente del efecto = «dos transeúntes»). Con esta información podemos generar una plantilla con los siguientes datos (omitimos los campos vacíos):

TYPE OF INCIDENT: MURDER
PERPETRATOR: ETA
OBJECTIVE: CIVILIAN: DOS TRANSEUNTES

Es importante destacar que se habría extraído la misma información si la oración hubiera sido *ETA mató a dos transeúntes en un atentado»* o *ETA perpetró un atentado en el que murieron dos*

[3] El tema fue propuesto por las Agencias que financian el proyecto.

transeúntes. Una de las ventajas de este tipo de modelos semánticos es que permite hacer un uso combinado de la información sintáctica y semántica.

Parámetros: el sistema permite trabajar con diferentes parámetros:

* *selection*: la información semántica está presente durante el análisis sintáctico, lo que permite usar restricciones seleccionales para mejorar y reducir considerablemente el número de análisis.

* para preservar la «robustez» del sistema contamos con la rutina *longest* que en el caso de no haber conseguido un análisis completo de toda la oración toma el análisis más largo que haya producido. También se puede limitar el número de *edges* que produce el parser, para que el tiempo de análisis no se prolongue más de un determinado punto.

* hay ciertas rutinas que reducen el número de analisis. Es importante tener en cuenta que la longitud media de las oraciones es de 30 palabras, que incluyen además varias estructuras incrustradas y múltiples modicadores. Por ejemplo, la rutina *closest* está pensada para establecer preferencias en la adjunción de modificadores. Las preferencias son diferentes para cada lengua. La rutina *subsume* también reduce el número de estructuras que se construyen durante el proceso de análisis sintáctico.

Las plantillas: el objetivo final es rellenar los campos de una plantilla con la información extraída por el sistema. Al igual que en una base de datos, las características de los campos tienen que ser definidas de antemano. El factor más importante que hay que tener en cuenta son las necesidades de los usuarios (es decir, recuperar la información que les interesa). La Agencia que financia el proyecto sugirió los siguientes campos:

```
1. MESSAGE ID:    identificación del mensaje resumido.
2. TEMPLATE ID:
3. DATE OF INCIDENT:
4. TYPE OF INCIDENT: p.ej. MURDER, ATTACK, KIDNAPPING, ROBBERY...
5. CATEGORY OF INCIDENT: p.ej  TERRORISM, STATE-SPONSORED ...
6. PERPETRATOR:ID OF INDIV(S): p.ej.  nombres propios o pseudónimos
7. PERPETRATOR: ID OF ORG:  p.ej. nombre de la organización
8. PHYSICAL TARGET: ID(S): p.ej. nombres de edificios, vías
                           de comunicación, etc.
9. PHYSICAL TARGET: TOTAL NUM:
10. PHYSICAL TARGET: TYPE(S):  p.ej. EMBASSY, ROAD, BRIDGE....
11. HUMAN TARGET: ID(S):
12. HUMAN TARGET: TOTAL NUM:
13. HUMAN TARGET: TYPE(S):  p.ej.  CIVILIAN, MILITARY, GOVERNMENT...
14. INSTRUMENT: TYPE(S):  p.ej.  EXPLOSIVE, GUN, ...
15. LOCATION OF INCIDENT:
16. EFFECT ON PHYSICAL TARGET(S):  p.ej. DAMAGE, DESTRUCTION...
17. EFFECT ON HUMAN TARGET(S): p.ej.  DEATH, INJURY.
```

3 Diferencias entre las versiones inglesa y española

La versión española empezó a desarrollarse en Septiembre de 1991 sobre los trabajos previos en inglés, algunos de ellos con bastantes años de investigación. Por tanto, ambas versiones se encuentran en distintos grados de desarrollo. Sin embargo, aunque el sistema para el español está basado en el del inglés, ha sido necesario utilizar distintas aproximaciones para adaptarse a los fenómenos particulares del castellano. De estas diferencias hablaremos a continuación.

3.1 Diccionario y morfología

En la sección 2.1 mencionamos que unas macros generaban todas las formas flexionadas siguiendo modelos paradigmáticos. Esto permite que el diccionario contenga sólo las formas básicas y la información morfosintáctica asociada.

Las características morfológicas del inglés y del español nos muestran la primera diferencia: las macros morfo-léxicas del castellano son mucho más complejas, tanto en número de paradigmas

como en definición de las formas flexionadas. Para la creación de los modelos se ha seguido la descripción de Moreno Sandoval (1991).

3.2 Gramática

El formalismo de la Lingustic String Grammar está basado en la teoría lingüística de Zelig Harris. Durante más de 15 años ha sido utilizado por Sager y sus colegas para desarrollar una gramática del inglés con una cobertura sintáctica muy amplia (Sager 1981). Se basa en la idea de representar las oraciones como secuencias de símbolos o *cadenas lingüísticas*, utilizando reglas sintagmáticas independientes del contexto. Todas las cadenas elementales están compuestas de un *núcleo* y sus modificadores o *adjuntos* a la derecha y a la izquierda. Los adjuntos a su vez pueden ser cadenas elementales. En el nivel oracional hay básicamente tres cadenas elementales: SUJETO, VERBO, OBJETO. Toda la gramática inglesa está adaptada a este esquema. Por ejemplo, dentro de OBJETO hay una gran variedad de subtipos que abarcan desde cadenas nominales y preposicionales aisladas, pasando por todas las combinaciones posibles de subconstituyentes[4], hasta oraciones subordinadas de ascenso, completivas y adverbiales.

Este planteamiento SVO se adapta muy bien al inglés (lengua originaria para la que estaba pensado el formalismo) pero ha habido que incorporar algunas variantes para dar cuenta de los fenómenos del español. Podemos resumir estas diferencias de la siguiente forma:

1) se permiten distintos órdenes de constituyentes para las oraciones afirmativas. Concretamente, SVO, VSO, VOS y OVS. El aumentar el número de posibilidades combinatorias entre las cadenas elementales conlleva el establecer restricciones que impidan las combinaciones incorrectas. Por ejemplo, la elisión del sujeto (en nuestra terminología sujeto «nulo») produciría 3 análisis completamente equivalentes. Nos interesa, por tanto, restringir la posibilidad del sujeto nulo sólo a un orden, SVO, con lo que evitamos los análisis redundantes con VSO y VOS. Otro tanto ocurre con los objetos «nulos» de los verbos, es decir, cuando se omite el objeto en construcciones intransitivas como «corre el niño» o «come el niño», que se pueden analizar como VSO_{nulo}, $VO_{nulo}S$ o $O_{nulo}VS$. Nosotros hemos optado por permitir únicamente el primer análisis, VSO.

2) los clíticos rompen con el esquema de las tres cadenas elementales ya que el OBJETO se fracciona en dos alrededor del VERBO («el abuelo LES regaló UN PATINETE»). Además se complica por el hecho de que los clíticos pueden aparecer en oraciones de ascenso («el abuelo LES obligó A CANTAR UNA JOTA») donde son a la vez OBJETO de la principal y SUJETO de la subordinada. Los clíticos aportan información redundante en muchos casos: «el abuelo LES regaló un patinete A SUS VECINOS». Tampoco hay que olvidar que puede haber dos clíticos juntos, antes o después del verbo. En posición enclítica presentan el problema adicional de reconocer que a una palabra gráfica le corresponden dos o tres palabras léxicas. Todos estos problemas han sido tratados de distintas maneras, de acuerdo con su naturaleza. Por ejemplo, la separación del verbo y los clíticos se realiza en el *scanner*. En el caso de los clíticos enfáticos, una rutina comprueba en el «análisis regularizado» si existen objetos redundantes y elimina el clítico si su presencia es innecesaria. Finalmente, como la idea de un «componente discontinuo» que agrupe los clíticos y el resto de los objetos no es fácil de implementar en nuestro formalismo, hemos optado por crear una cadena especial CLITICOS, que permite establecer adecuadamente las distintas funciones en la forma regularizada que se produce al final del análisis sintáctico. La estructura básica queda ahora de la siguiente manera:

<SUJETO> <CLITICOS><VERBO><CLITICOS><OBJETO>.

3) La posibilidad de elisión del sujeto es una diferencia significativa con respecto a la gramática inglesa. Dado que la información sobre persona y número está presente siempre en el verbo, se utiliza

[4] Por ejemplo, la combinación de objeto directo e indirecto, o de completiva y objeto indirecto, etc.

una rutina (*unify*[5]) que permite copiar dichos rasgos en la posición del SUJETO en la estructura regularizada. De esta forma, la información está disponible en la parte discursiva para computar la resolución de las anáforas.

4) Hay tratamientos particulares para ciertas construcciones propias del castellano como las cláusulas con «se», los usos de los verbos auxiliares, los sintagmas de tiempo y fechas, las construcciones de lugar, etc.

3.3 Los modelos semánticos

Dado que se utiliza una aproximación de semántica léxica, para cada lengua es necesario crear unos modelos que tengan en cuenta las particularidades sintácticas de los elementos léxicos. Por tanto, para cada concepto relevante[6] existe un modelo que proyecta la estructura sintáctica de los argumentos en los papeles temáticos asociados. Podemos dar cuenta así de que los sujetos de las construcciones con «se» son los pacientes del suceso, o establecer la relación conceptual que existe entre «herir» y «resultar herido». En Olmeda y Moreno(1992) se proporciona un ejemplo de la implementación de estos modelos.

4 Un ejemplo

Presentaremos a continuación un ejemplo del grado actual de desarrollo de la versión española. El texto escogido es un parte de la agencia EFE. Lo mostramos tal y como el sistema es capaz de leerlo; es decir, en este momento el programa no reconoce los caracteres nacionales del castellano como las vocales acentuadas y las «ñ». Utilizamos en su lugar la convención:

á = a1 é = e1 í = i1 ó = o1 ú = u1 ñ = n1

El texto escogido es significativo por cuanto tiene complejidad y variedad de construcciones sintácticas (la longitud media de las oraciones es de 34 palabras) al tiempo que su contenido ejemplifica el dominio temático de la aplicación.

SAN SALVADOR, 12 JAN 90 (ACAN-EFE) -- [TEXT] El gobierno salvadoren1o ha manifestado su condena por la desaparicio1n del lilder socialista Helctor Oqueli Colindres esta manlana en Guatemala.

El Ministro de Informacio1n Mauricio Sandoval dijo a ACAN-EFE que «era lamentable», anladiendo que el gobierno salvadoren1o haral los esfuerzos necesarios a travels de los canales diplomalticos para obtener mals informacio1n sobre el caso.

De acuerdo con las fuentes de su partido Helctor Oqueli Colindres, secretario general del Movimiento Nacional Revolucionario (MNR), desapareciol hoy en Guatemala cuando el vehilculo en el que viajaba fue interceptado por un grupo de hombres fuertemente armados que vestilan atuendos civiles.

---

[5] Unify se utiliza también para la comprobación de los valores de los rasgos (por ejemplo, de la concordancia). En la práctica es equivalente a la operación de «unificación» propia de los formalismos expuestos en Shieber(1986), con la diferencia de que en nuestro caso se hace de forma procedural en lugar de declarativa, y además es más «débil» que la unificación pues no realiza el ligado de variables.

[6] Ya que crear un modelo semántico general con todos los conceptos de una lengua natural es una tarea inviable, nos limitamos a codificar los que presentan una significación especial para el dominio que tratamos.

El portavoz del MNR dijo que Oqueli Colindres fue «secuestrado» cuando se dirigía al aeropuerto de la ciudad de Guatemala para tomar un vuelo hacia Nicaragua, donde iba a reunirse con la delegación socialista internacional que se encuentra en dicho país observando la campaña electoral.

El portavoz del MNR dijo también que la policía guatemalteca encontró, sin sus ocupantes, el vehículo en el que viajaba Oqueli con un miembro del Partido Socialdemócrata de Guatemala.

Texto 1.

Como mencionamos al principio de artículo, para evaluar los resultados necesitamos compararlos con los producidos manualmente. Un especialista en el tema rellenó la plantilla de la siguiente manera[7]:

```
1. MESSAGE ID:                                  DEV-MUC3-0026
2. TEMPLATE ID:                                 1
3. DATE OF INCIDENT:                            12 JAN 90
4. TYPE OF INCIDENT:                            KIDNAPPING
5. CATEGORY OF INCIDENT:                        -
6. PERPETRATOR:ID OF INDIV(S):                  «UN GRUPO DE HOMBRES
                                                     FUERTEMENTE ARMADOS»
7. PERPETRATOR: ID OF ORG:                      -
8. PHYSICAL TARGET: ID(S):                      *
9. PHYSICAL TARGET: TOTAL NUM:                  *
10. PHYSICAL TARGET: TYPE(S):
11.HUMAN TARGET:ID(S):                          «HÉCTOR OQUELI COLINDRES»
              («SECRETARIO
              GENERAL DEL MOVIMIENTO NACIONAL REVOLUCIONARIO»)
12. HUMAN TARGET: TOTAL NUM:                    1
13. HUMAN TARGET: TYPE(S):                      POLITICAL FIGURE: «HÉCTOR
                                                     OQUELI  COLINDRES»
14. INSTRUMENT: TYPE(S):                        *
15. LOCATION OF INCIDENT:                       GUATEMALA: CIUDAD DE GUATEMALA
16. EFFECT ON PHYSICAL TARGET(S):               *
17. EFFECT ON HUMAN TARGET(S):                  -
```

La plantilla generada por nuestro sistema es exactamente igual excepto en el campo de «localización del incidente»: PROTEUS no es capaz de reconocer que el secuestro tuvo lugar en Ciudad de Guatemala, aunque sí consiguió recuperar el país, Guatemala. En este caso hemos obtenido un 100% de recall y un 94% de precisión[8].

Para conocer cómo afecta a los resultados el hecho de poner algún tipo de límite al tiempo de procesamiento hicimos el siguiente experimento. Limitamos el número de *edges* a 25000 y obtuvimos que el PERPETRATOR del secuestro fue «UN GRUPO». Pusimos un nuevo límite en 20000 y esta

---

[7] Los conceptos claves (la fecha, el tipo de incidente, el tipo de objetivo humano) están en inglés pues utilizamos el generador de plantillas de ese idioma. El asterisco significa que la información de ese campo no es pertinente para el tipo de SUCESO que se está analizando. El guión significa que esa información es pertinente, pero o bien no está en el texto o bien no se ha podido extraer.

[8] Es decir, si consideramos que la localización ha sido rellenada de forma correcta parcialmente, entonces tenemos 8.5 *slots* correctos sobre 9. Si contamos como equivocada la localización entonces 8/9 = 88%

vez el sistema no fue capaz de reconocer el agente. Esto se explica porque la oración principal del mensaje (i.e. el párrafo tercero) tiene una longitud de 42 palabras y el agente se encuentra al final de la oración. Si ponemos un límite de tiempo el sistema no llega a procesar hasta el final y por lo tanto perdemos ese elemento de información. Como contrapartida el tiempo real de procesamiento de todo el mensaje (171 palabras repartidas entre 5 oraciones) se ha reducido casi a la mitad: aproximadamente de 18 minutos a 10 minutos, en una Sun Sparc2 con 34 Mg RAM y 50 Mg swap.

Los experimentos realizados con la versión inglesa con 100 mensajes similares al del ejemplo han demostrado que es preferible perder algo de precisión a cambio de ganar mucho tiempo de procesamiento, y por tanto es más eficiente en la actualidad poner algún tipo de limitación temporal.

5 Conclusiones

La evaluación más reciente de la versión inglesa sobre un corpus bastante representativo sobre terrorismo ha dado unos resultados del 44% de recall y 57% de precisión (Grishman et al. 1991a)[9]. Dichos resultados están lejos de ser satisfactorios pero permiten albergar esperanzas sobre el futuro de la extracción automática de información, en el sentido de que los textos complejos empiezan ya a ser analizados e interpretados con cierta eficacia.

Con respecto a PROTEUS, la implementación de la versión española esta permitiendo comprobar la robustez y portabilidad del sistema así como desarrollar nuevas herramientas y mejorar las existentes para convertirlo en un auténtica aplicación multilingüe. En gran medida, el hecho de que se disponga de una versión española en menos de un año es una prueba de la madurez de todo el sistema.

6 Agradecimientos

El proyecto PROTEUS está financiado por la Defense Advanced Research Projects Agency con la beca N00014-90-J-1851 de la Office of Naval Research, y por la National Science Foundation con la beca IRI-89-02304.

La investigación de Antonio Moreno Sandoval está siendo subvencionada por una beca postdoctoral M.E.C.-Fulbright.

7 Referencias

Grishman, R. (1991): «Information Extraction from Natural Language Text». PROTEUS Project Memorandum num. 47, Department of Computer Science, New York University.

Grishman, R., Sterling, J. y Macleod, C. (1991a): «New York University PROTEUS System: MUC-3 Results and Analysis», en *Proceedings of the Third Message Understanding Conference-3.* 95-98.

Grishman, R., Sterling, J. y Macleod, C. (1991b): «Description of the PROTEUS system as used for MUC-3», en *Proceedings of the Message Understanding Conference-3.* 183-190.

Mauldin, M. (1991): «Retrieval Performance in FERRET: A Conceptual Information Retrieval System», en *Proc. of ACM SIGIR-91.* Chicago, Oct. 13-16, pp. 347-355.

Moreno Sandoval, A. (1991): *Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español.* Tesis doctoral, Universidad Autónoma de Madrid.

[9] Para comparar estos resultados con los obtenidos por otros sistemas que han trabajado sobre el mismo corpus y la misma plantilla de respuesta consúltese *Proceedings of the Third Message Understanding Conference (MUC-3)*, donde se recogen los resultados de una evaluación exhaustiva realizada por el Naval Ocean Systems Center. Se analizan los sistemas de quince equipos de investigación americanos que representan una amplia variedad de técnicas de interpretación textual (palabras-clave, estadística, procesamiento lingüístico, etc.)

Olmeda, C. y Moreno, A. (1992): «El tratamiento semántico en un sistema automático de extracción de información», PROTEUS Project Memorandum num. 50, Department of Computer Science, New York University.

Proceedings of the Message Understanding Conference-3. San Mateo, CA: Morgan Kaufmann, 1991. ISBN 1-55860-236-4.

Sager, N. (1981): *Natural Language Information Processing: A Computer Grammar of English and Its Applications.* Reading, MA: Addison-Wesley.

Shieber, S. (1986): *An Introduction to Unification-based Approaches to Grammar.* Stanford: CSLI.

Sundheim (1991): «Overview of the Third Message Understanding Evaluation and Conference», en *Proceedings of MUC-3.*