Reconocimiento de entidades en el sistema EXIT

Rafael Muñoz, Andrés Montoyo, Fernando Llopis, Armando Suárez.
Grupo de Programación Lógica y Sistemas de Información.
Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.
Tel.: 96-590-34-00. Fax:96-590-93-26
{rafael, montoyo, llopis, armando}@dlsi.ua.es

Resumen

EXIT es un sistema de extracción de información en el dominio de las escrituras notariales de compra-venta. En este trabajo se presenta el mecanismo que se adopta para la identificación y reconocimiento de entidades en el sistema EXIT. Fundamentalmente las entidades que son objeto de extracción son las organizaciones y/o personas que intervienen en las operaciones de compra-venta, así como los inmuebles objetos de la transacción. Para la identificación y reconocimiento de estas entidades nos basamos en: a) una serie de disparadores (palabras específicas que introducen o forman parte de las entidades); b) un conjunto de reglas específicas que determina la estructura de una entidad; c) heurísticas para la desambiguación. Además se presenta una visión genérica del sistema EXIT presentado en [LLOP98].

1. Introducción.

Al estilo de otros sistemas de extracción de información como LaSIE [GAI95] o TURBIO [TUR96], EXIT es un sistema de extracción de información que trabaja con un dominio restringido, en concreto, en el dominio de las escrituras notariales de compra-venta. Fundamentalmente la información a extraer comprende las diferentes entidades y los inmuebles que aparecen en dichas escrituras. El sistema no solo deberá reconocer dichas entidades sino además solucionar los múltiples problemas de referencias anafóricas que se presenten, con la complejidad añadida que dichas referencias implican a entidades que no se hallan en la misma frase o párrafo, sino en párrafos especificados con bastante anterioridad. Para la resolución de estos problemas de anáforas utilizamos los métodos descritos en [FER97a][FER97b] .

El objetivo de EXIT, al igual que la mayoría de los sistemas de extracción [CRA96][MAS], es el de generar una serie de plantillas en la cual se almacene la información relevante extraída de las escrituras. Las plantillas que genera EXIT son:

- Implicados u otorgantes: Contiene la información referente a las personas y/u organizaciones que intervienen en la operación de compra-venta. Tendremos tantas plantillas de este tipo como personas y/u organizaciones sean identificadas en el texto. La información que contienen estas plantillas será distinta en función que se refieran a personas u organizaciones. Los datos que contiene las plantillas de personas serán nombre completo, dirección, DNI o NIF, etc. Los datos de las plantillas de organizaciones será nombre, dirección, CIF, etc.
- Inmueble: Contiene información del inmueble descrito en la operación, como tipo de inmueble, características, dirección, etc..
- Operación: Contiene diversa información. Como es el nombre del notario que da fe de la operación, lugar donde se produce, fecha, etc.

2. Arquitectura del Sistema EXIT.

La figura 1 muestra la arquitectura del sistema EXIT, en la cual podemos distinguir las siguientes fases o etapas, que a continuación se describen brevemente:

1. *Tokenización y etiquetado*. La información de entrada al sistema será una serie de escrituras de compra-venta que se encuentran en formato electrónico. El tokenizador procesa cada una de las escrituras identificando palabras y asignando a cada uno de ellos un identificador que permanece durante todo el proceso. Paralelamente el etiquetador procesa cada palabra asignándole las categorías gramaticales posibles que pueda tomar dicha palabra. En esta etapa se utiliza un diccionario de propósito general o un módulo probabilístico para la obtención de las categorías. Al finalizar esta etapa obtendremos todos las palabras con todas las categorías gramaticales que puedan ser. No obstante pueden existir tokens para los cuales no se encuentren en el diccionario de propósito general, en cuyo caso se etiquetará como palabra desconocida.
2. *Tokenizador específico*. La salida de la fase anterior se pasa a un módulo para la identificación de entidades. Para EXIT contamos con los siguientes diccionarios:
   - Diccionario de nombres (de 4337 entradas)
   - Diccionario de apellidos (de 4657 entradas)
   - Diccionario de localidades (de 53000 entradas)
   - Diccionario de actividades

   Este módulo añadirá una etiqueta a las palabras que hayan sido identificadas como nombre propio o palabra desconocida y aparezca en alguno de estos diccionarios, con una etiqueta específica en función del diccionario en el que hayan sido encontrados.
3. *Reconocedor de entidades*. Esta etapa o módulo se apoya en una gramática específica para identificar entidades o partes de entidades, como es la dirección. En este artículo se presenta esta fase de una forma más detallada,
4. *Analizador*. En esta fase se realiza un análisis sintáctico de las frases apoyándose en una gramática y en una ontología de rasgos, obteniéndose una serie de estructuras sintácticas.
5. *Resolución de correferencias*. A partir de las estructuras generadas en la etapa anterior se resuelven las correferencias existentes en el texto y se generan unas estructuras que no tengan correferencias.
6. *Interpretación semántica*. Esta fase obtiene una forma lógica independiente del contexto a partir de la información generada en la fase anterior.
7. *Interpretación del discurso*. Por último se realiza una interpretación del discurso basándose en un modelo del mundo que se obtiene de una base de hechos y una serie de reglas de inferencia.

3. Reconocimiento de entidades en EXIT.

El objetivo de esta etapa consiste en estudiar si una o más palabras forman una entidad e identificar el tipo de la misma. Es decir, comprobar si una serie de palabras identifican una entidad, como puede ser una persona u organización, una dirección, una localidad, etc.

Como entrada de esta etapa disponemos de un conjunto de palabras con su categorías gramaticales (token), como consecuencia del proceso de las etapas anteriores, así como una gramática específica de entidades. Esta gramática de entidades contiene una serie de reglas gramaticales correspondiente a la estructura de las entidades.

Las características fundamentales que deben cumplir las palabras que van a formar parte de las entidades son:

a) palabras en mayúsculas que no sean inicio de frase,

b) palabras que no han podido ser catalogadas con una etiqueta gramatical,
c) una serie de palabras (disparadores) que suelen acompañar a algunas entidades buscadas. Por ejemplo la palabra *Don*, precede a una entidad persona, o la palabra *avenida* suele preceder a una dirección. Cuando aparezcan dichos disparadores se corresponderá a una serie de reglas.

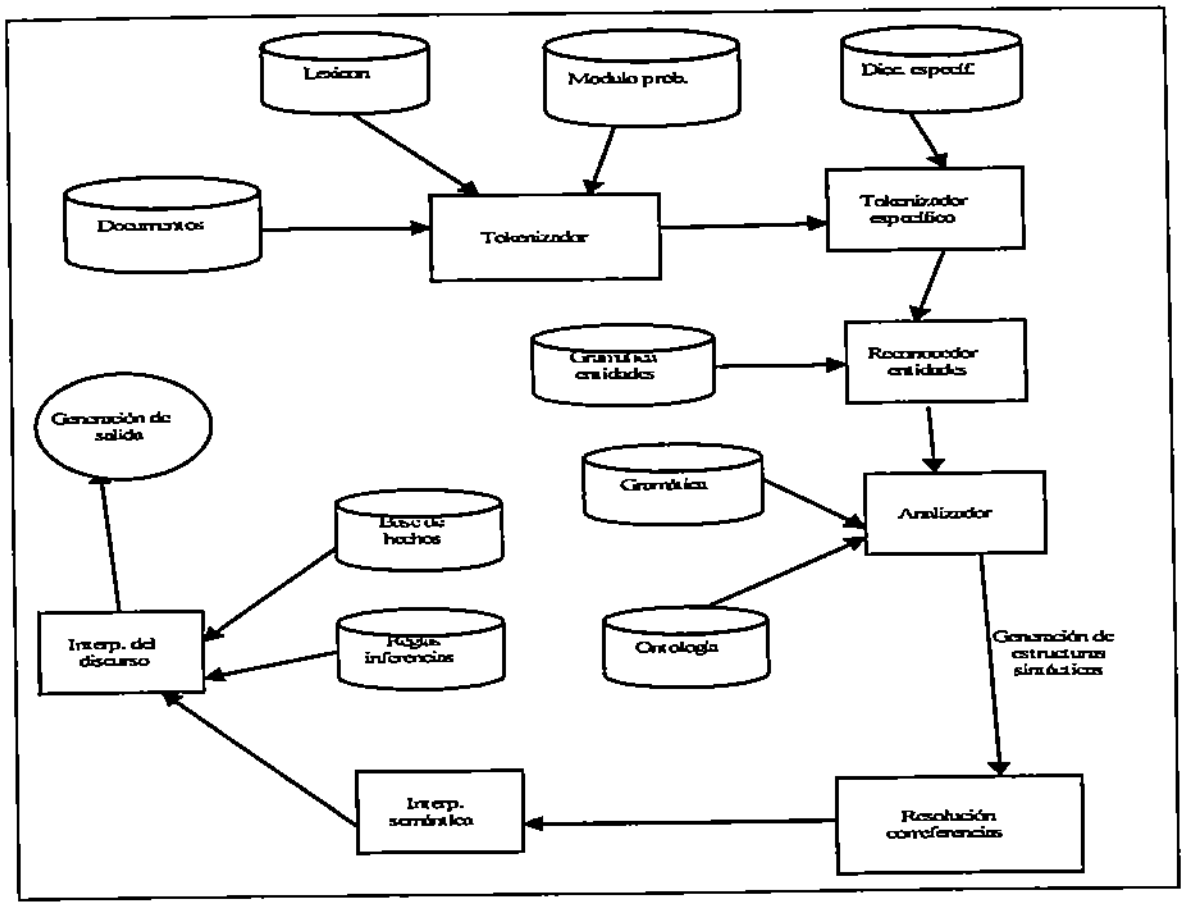

Figura 1. Arquitectura del sistema EXIT

La dificultad que conlleva esta etapa o fase es identificar el inicio y fin de la entidad. Uno de los problemas que nos vamos a encontrar es la aparición de preposiciones, artículos y conjunciones, que pueden aparecer en minúsculas en medio de la entidad que queremos reconocer. Por ejemplo la preposición "de" y el artículo "las" en el nombre *Francisco García de las Heras* o en la localidad *Cabo de las Huertas.*

A continuación vamos a ver el conjunto de reglas específicas para cada tipo de entidad a reconocer que forman la gramática de entidades anteriormente mencionada, así como una serie de disparadores que nos indicarán el tipo de entidad que estamos identificando.

La nomenclatura en la que se expresan las reglas es la utilizada en los diccionarios de datos de los sistemas de información. Las palabras que aparecen entre paréntesis significa que son opcionales, es decir, que podrán aparecer o no en el patrón. Las palabras que aparecen entre corchetes significa que aparecerá en la regla únicamente uno de ellos. Las palabras que aparecen entre llaves significan que habrá una repetición de una o más palabras en la regla.

3.1 Etapas en la identificación y reconocimiento de entidades.

Para la identificación y reconocimiento de entidades diferenciamos cuatro etapas fundamentales, que son:

1. Identificación de los disparadores.
2. Reconocimiento de nombres, organización ó dirección
3. Desambiguación
4. Reconocimiento de entidades

A continuación vamos a describir de forma más detallada los módulos de reconocimiento de personas, organizaciones ,y direcciones y localidades.

3.1.1 Módulo de reconocimiento de personas.

Las entidades persona que aparecen en los textos, están escritas en mayúsculas, bien las iniciales o bien de forma completa, y suelen estar precedidas por algún disparador. Además las palabras que forman parte de la entidad persona suelen hallarse en los diccionarios de nombres y apellidos que es consultado por el sistema, y por tanto hay una alta probabilidad de que hallan sido identificados como nombres de persona o apellido. En caso de que no aparezca en ninguno de los diccionarios estará identificada como palabra desconocida (pdes).

Otro problema que nos encontramos es la aparición de las preposiciones *"de"* y *"del"*, de los artículos *"las"* y *"la"* en el apellido, como hemos comentado anteriormente. También puede aparecer la conjunción "*y*" que lo que hace es enlazar dos apellidos, pero no forma parte de ninguno de ellos. La desambiguación de nombres solucionará este problema viendo si esa conjunción *"y"* enlaza dos entidades o dos apellidos de una sola entidad.

Reglas de entidades personas:

- entidad → ["D." | "DON" | "DOÑA" | "SEÑOR" | "SEÑORA" | "SR." | "SRA."] + *persona*
- *persona* → nombre + (nombre) + *apellido* + (y) + (*apellido)*
- *nombre* → [npNOM | pdes]
- *apellido*→ [pdes | npAPE] + (prep + (art) + [pdes | npAPE])
- *apellido* → prep + (art) + [pdes | npAPE]

3.1.2 Módulo de reconocimiento de organizaciones.

Del mismo modo que para el reconocimiento de personas las palabras que van a formar parte de la entidad organización deben empezar por mayúsculas. También puede ocurrir que todos los caracteres que forman la entidad vayan en mayúsculas.

Reglas de entidades organización:

- *organización* → *nombre* + [SL | S.L. | SA | S.A. | CO. | Cia. | SAD | S.A.D. | Inc. | S. Coop. | Coop. V. | C. V. | S.C.V.]
- *organización* → [S.C.D. | S. Coop. | Coop. V. | C.V. | S.C.V.] + *nombre*
- *nombre* → {palabras}
- Hay una serie de disparadores que aparecerán en unas reglas generales para el reconocimiento de organizaciones. Estos disparadores son los que aparecen en comillas en las reglas que se presentan a continuación.
  - "la entidad mercantil" + *organización*
  - "la entidad" + *organización*
  - "la entidad acreedora" + *organización*
  - "la sociedad" + *organización*
  - "la compañía" + *organización*
  - "la empresa" + *organización*
  - "la multinacional" + *organización*
  - "la institución" + *organización*
  - "apoderado de" + *organización*
  - Hay muchas organizaciones que vienen identificadas por la actividad que desarrollan, por ejemplo Bar-Restaurante Costablanca, Clínica San Carlos, Club De Tenis De Alicante. Por lo tanto en Exit se dispone de una lista de aquellas actividades más frecuentes

ordenadas por orden alfabético. Veamos el siguiente ejemplo: Clínica San Carlos. Como la palabra clínica aparece en el diccionario de actividades y aparece acompañado de unas palabras en mayúsculas identificamos todo como una organización.

3.1.3 Módulo de reconocimiento de direcciones y localidades

Para el reconocimiento de direcciones en un texto se utilizan también una serie de disparadores que aparecerán en unas reglas generales. Estos disparadores son los que aparecen entrecomillados en las reglas que a continuación se presentan. Los disparadores serán verbos o sintagmas nominales que introducen una estructura fija en los textos de compra-venta. Por ejemplo si queremos identificar la dirección en la frase *D. Javier García Pérez vecino de Alicante, calle Alfonso el Sabio, número 2, bloque 1, 3º izq.* El sintagma nominal *"vecino"* seguido de un sintagma preposicional *"de ...."* actúa de disparador de un tipo de patrón concreto.

Reglas de direcciones y localidades:

- "vecino de" + *localidad* + ", " + *dirección*
- "con residencia en" + *localidad*
- "domiciliada en" + *localidad* + "," + *dirección*
- "en termino de" + *localidad*
- "con domicilio en" + *dirección*
- "sito en" + *dirección* + "," + *localidad*
- *localidad* → { [nprop | prep | art] }
- *dirección* → *tipo_vía* + { [nprop | prep | art] } + "," + *número* + "," + *tipo_inmueble*
- *tipo_vía* → ["alameda" | "avenida" | "calle" | "camino" | "carrera" | "carretera" | "cuesta" | "colonia" | "glorieta" | "partida" | "pasaje" | "plaza" | "polígono" | "paseo" | "prolongación" | "ronda" | "travesía" | "urbanización"]
- *tipo_inmueble* → (["apartamento" | "chalet" | "bungalow" | "edificio" | "caserío" | "nave" | "local"]) + ((*bloque*) + (*portal*) + (*escalera*) + (*piso*) +(*puerta*))
- *bloque* → "bloque" + cardinal + ","
- *portal* → cardinal + "portal" + ","
- *portal* → "portal" + cardinal + ","
- *número* → (["número" | "num." | "nº." | "n."]) + cardinal
- *número* → "s/n"
- *escalera* → ["escalera" | "esc."] + cardinal + ","
- *escalera* → cardinal + ["escalera" | "esc."] + ","
- *piso* → ("piso") + cardinal
- *piso* → cardinal + ("piso")
- *puerta* → ("letra") + carácter
- *puerta* → ["izquierda" | "izq." | "derecha" | "dcha."]

3.2 Desambiguación de nombres de entidades

En el proceso anterior puede darse el caso que se hayan identificado dos o más entidades como una sola. Por ejemplo I.B.M. y Microsoft. Consideraremos como operadores ambiguos, las conjunciones y las preposiciones. El sistema EXIT aplica los siguientes pasos para resolver la ambigüedad estructural (según [WAC]):

- Se busca la posición del operador ambiguo y se evalúan por separado las subcadenas de la derecha y de la izquierda.
- Se aplican una serie de heurísticas a las subcadenas para saber si se trata de una única entidad o por el contrario forman más de una entidad.

A continuación se exponen las heurísticas que se aplicarán para resolver la ambigüedad:

1. Siempre que aparezcan disparadores de personas u organizaciones en las dos subcadenas se dividen en dos entidades. Son disparadores de personas los anteriormente descritos como son DON, SEÑOR, etc. Y disparadores de entidades son S.A., S.L., etc. Ejemplo:

- *Telefónica S.A. e Internet Sistemas S.L.*
  Se identifica dos subcadenas. La subcadena *Telefonica S.A.* y la subcadena *Internet Sistema S.L.* Como las dos cadenas tienen disparadores se dividen en dos entidades.

2. Si en alguna de las dos subcadenas no se encuentran ninguno de los disparadores. Pueden ocurrir los siguientes casos:

2.1. La cadena de la izquierda contenga el disparador, con lo cuál ya tenemos una entidad, por lo que la subcadena de la derecha debe ser otra entidad distinta. Ejemplo:

- *MECEMSA S.A. y Vicente García*
  La subcadena izquierda *MECEMSA S.A.* contien el disparador *S.A.* por lo que sabemos que son dos entidades.

2.2. La cadena de la izquierda no contenga ningún disparador y la de la derecha sí. Como la parte de la derecha tiene un disparador sabemos que la subcadena de la derecha es una entidad. Ahora hay que ver si la subcadena de la izquierda es una entidad propia o forma parte del nombre de la entidad de la subcadena de la derecha. Si en la subcadena de la izquierda se identifica alguna actividad entonces forma una entidad propia, en caso contrario forma parte del nombre de la entidad de la derecha. Pueden existir casos en los que sea necesario realizar una posterior desambiguación semántica. Ejemplos:

- *Clínica San Carlos y Hercules S.A.D.*
  Como la subcadena *Hercules S.A.D.* tiene un disparador entonces decimos que una es una entidad. Ahora habrá que estudiar la subcadena de la izquierda, *Clínica San Carlos*. Como podemos ver tiene una actividad, "clínica", por lo tanto forma una entidad propia.
- *Ortiz e Hijos S.A.*
  Como la subcadena de la derecha *Hijos S.A.* tiene un disparador entonces decimos que una es una entidad. Ahora habrá que estudiar la subcadena de la izquierda, *Ortiz*. Como podemos ver no tiene ninguna actividad, por lo tanto no forma una entidad propia y la subcadena izquierda es parte de la entidad. Es decir, tenemos una sola entidad que es *Ortiz e Hijos S.A.*
- *Colegio Jesús y María S.A.*
  Como la subcadena de la derecha *María S.A.* tiene un disparador entonces decimos que una es una entidad. Ahora habrá que estudiar la subcadena de la izquierda, *Colegio Jesús*. Como podemos ver tiene una actividad, colegio, por lo tanto forma una entidad propia. Es decir, tenemos dos entidades, lo cual es incorrecto, por lo que hará falta realizar la desambiguación semántica anteriormente mencionada.

3. Si ninguna de las dos subcadenas dispone de disparadores se aplican las mismas heurísticas pero teniendo en cuenta la actividad en lugar de los disparadores. Ejemplo:

- *Arjame Mociones y Construcciones*
  La subcadena izquierda contiene una actividad y un sintagma nominal, por lo tanto se identificaría como una entidad. La subcadena derecha contiene una actividad pero no le acompaña un sintagma nominal, por lo que no forma una entidad propia. Sino que la cadena completa es una única entidad.

4. Si hay un acrónimo a la izquierda o derecha del operador ambiguo siempre se divide en dos entidades. Ejemplos:

- *CAM y Banco de España*
  Como CAM es un acrónimo de Caja de Ahorros del Mediterráneo entonces se divide en dos entidades
  Otro ejemplo sería, *Laboratorios de IBM y Microsoft.*

4. Conclusiones.

En este trabajo se presenta un mecanismo para la identificación y reconocimiento de entidades para el sistema EXIT basado en un conjunto de reglas que incluyen una serie de disparadores que nos

permiten identificar las entidades; estos disparadores que son palabras específicas que introducen o forman parte de una entidad han sido extraídos de forma manual a partir de un corpus de escrituras de compra-venta.

Además, se presenta un mecanismo para la desambiguación de nombres necesario para el reconocimiento de las entidades.

En estos momentos estamos desarrollando los módulos presentados anteriormente, utilizando PROLOG desde dos puntos de vista: gramáticas regulares y gramáticas lógicas.

Pretendemos comprobar la eficiencia de este módulo midiendo la precisión y cobertura utilizando un corpus de escrituras (alrededor de 200 escrituras) y compararlo con otros métodos existentes.

5. Bibliografía.

[COW96] Jim Cowie, Wendy Lehnert. "*Information Extration*". Comunications of ACM 1996.

[CRA96] Malcon Crawford. "*Information Extraction*". The University of Sheffield. http://www.dcs.shef.ac.uk/research/groups/nlp/extraction/. 1996.

[FER97a] Ferrández, A.; Moreno, L.; Palomar, M.; Peral, J. "*Un método de resolución de la anáfora discursiva mediante la unificación*". Congr. Asociación Española Para la Inteligencia Artificial. 1997.

[FER97b] Ferrández, A.; Palomar, M.; Moreno, L. "*Slot Unificacion Grammar and anaphor resolution*". Recent Advances in Natural Language Resolution.1997

[GAI95] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, Y. Wilks. "*Description of the LaSIE system as used for MUC-6*". Proceedings of the Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann, 1995, pp. 207-220.

[GRIS95] Ralph Grishman. "*MUC-6 - Documment*". New York University. 1995. http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

[LEH] Wendy Lehnert. "*Information Extraction*". University of Massachussetts. http://www.cs.buffalo.edu/~gherrera/inf-ext.html

[LLOP98] F. Llopis et. al. "*EXIT: Propuesta de un sistema de extracción de información de textos notariales*". Novática nº 133 (mayo–junio 1998).

[MAS] Natural Language Processing Laboratory, University of Massachusetts. "*Information Extraction*". http://www-nlp.cs.umass.edu/~nlgroup/

[MILL93] G. Miller et al. "*Five papers on WordNet*". Technical reports CSL Report 43 Cognitive Sciencè Laboratory. Princeton University 1993.

[MOR92] L.Moreno, M.Palomar, F.Andrés. "*Incorporar Restricciones Semánticas al Análisis Sintáctico*". Procesamiento de Lenguaje Natural nº12. 1992.

[TUR96] Jordi Turmo Borrás "*TURBIO: Sistema de extracción de información a partir de textos estructurados*". DLSI. UPC. 1996

[VOS97] Piek Vossen, Pedro Díez-Orzas, Wim Peters "*Multilingual design of EuroWordNet*". 1997 Proceedings of ACL/EACL97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources

[WAC] N. Wacholder, Y Ravin, M. Choi. "Disambiguation of Proper Names in Text"