Extracción automática de información de fuentes documentales históricas mediante técnicas de procesamiento de lenguaje natural

Por

Antonio Calvo Cuenca

Introducción

Planteamiento general del problema

La actividad investigadora conlleva un conjunto de tareas y actividades encaminadas a la elaboración de trabajos de investigación que se plasman en forma de ponencias a congresos, artículos, tesis doctorales, monografías, etc. Desde que se plantea un problema hasta que se obtiene la publicación de un trabajo donde se expresan nuevas teorías se llevan a cabo toda una serie de tareas guiadas por una metodología en la que se emplean diferentes técnicas, herramientas y recursos. Con la ayuda de esta metodología se pretende obtener trabajos de buena calidad con la máxima eficiencia. Este trabajo plantea nuevas propuestas sobre técnicas y herramientas en el dominio de la investigación histórica.

Arquitectura del sistema propuesto

Para el desarrollo de la aplicación se ha utilizado como arquitectura de referencia la arquitectura Modelo de la Aplicación, Vista, Controlador (MVC). En este trabajo sólo abordaremos el componente del modelo de la aplicación, describiendo las principales tareas que se llevan a cabo, así como el conocimiento del dominio de la aplicación. Las tareas corresponderán a procesos que manipulan elementos del conocimiento del dominio como las fuentes docu-

mentales, hacen uso de bases de conocimiento del dominio y dan como resultado otros elementos del dominio como tablas que representan estructuras de síntesis o documentos de síntesis.

Preprocesamiento I: normalización

La figura 1 muestra parte de la arquitectura del modelo de la aplicación dentro de la arquitectura global del sistema. Un conjunto de documentos no normalizados es sometido a un primer preprocesamiento, normalizando la codificación de los caracteres y buscando una representación en forma de texto uniforme. La salida de este preprocesamiento es un corpus de documentos xml donde se ha definido un elemento texto y dentro de él se han definido un conjunto de elementos denominados p, que representan párrafos del texto.

Preprocesamiento I: clasificación de párrafos

Estos párrafos del texto se clasifican temática y estructuralmente a partir de un conocimiento que ha sido declarado en la ontología general a la que hemos llamado ontología histórica. Esta clasificación queda reflejada en los atributos del elemento p como puede observarse en la figura 2. Se han utilizado varios métodos para llevar a cabo esta clasificación (métodos intensivos en conocimiento –método de la poda– y métodos estadísticos –bayesianos y basados en árboles de decisión–). Disponer de un corpus de documentos con los párrafos del texto clasificados facilita la labor de recuperación de información.

Extracción automática de nombres de entidades y de relaciones

Un segundo módulo de preprocesamiento, ver figura 1, está dedicado a la extracción de nombres de entidades. El objetivo es que a partir del conjunto de documentos normalizados y clasificados se pueda identificar y extraer de forma automática un conjunto de entidades y relaciones de interés para el investigador, de forma que permita obtener en un nuevo documento de síntesis una representación de las fuentes documentales donde, además de incluir los propios documentos, podamos también incluir estructuras de síntesis (tablas de hombres, mujeres, instituciones y lugares), índices onomástico, toponímico y temático. El salto cualitativo es importante pues contar con este documento de síntesis facilita la localización de todas estas entidades. Lograr este objetivo no es fácil, pues esta tarea incluye además de la identificación de las entidades y relaciones, la resolución de problemas de coo-referencia y la resolución de aná-

foras en el texto mediante técnicas de procesamiento de lenguaje natural. Este preprocesamiento contiene las subtareas que se describen a continuación.

Segmentación

La tarea de segmentación consiste en identificar y aislar todos los elementos léxicos del texto. La figura 3 muestra el texto original y debajo el texto segmentado. Puede observarse cómo se han aislado los signos de puntuación de las palabras del texto. A cada uno de estos componentes léxicos los denominaremos token.

Etiquetado

El etiquetado del texto consiste en asignar a cada componente léxico obtenido de la tarea anterior una etiqueta que representa la categoría léxica del término en la oración. Las categorías léxicas se corresponden a nombres comunes, nombres propios, adjetivos, artículos determinantes, etc. El analizador morfológico para el castellano desarrollado utiliza un conjunto de etiquetas para representar el aspecto morfológico de las palabras. Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas[1]. Para el etiquetado de los componentes léxicos (tokens) hacemos uso de una técnica basada en expresiones regulares[2].

El programa de etiquetado leerá e interpretará este fichero para crear y aplicar el proceso algorítmico de etiquetado. La figura 4 muestra un fragmento de este fichero de reglas de etiquetado. La tercera parte de la figura 3 muestra el resultado de aplicar la tarea de etiquetado al resultado previo de segmentación.

Análisis sintáctico parcial

El siguiente paso en la arquitectura propuesta en la figura 1 es el análisis sintáctico parcial (ASP). En este trabajo hemos utilizado una técnica lingüística basada en gramáticas. Se ha creado un fichero de texto que contiene la gramática como se observa en la figura 4. El procesador lee este fichero, construye con

[1] http://www.ilc.cnr.it/EAGLES96/intro.html.

[2] BIRD, S., KLEIN, E. and LOPER, E., *Natural Language Process-ing with Python*, O'Reilly Media, 2009. cap. 5.

él un reconocedor de nombres de entidades que aplica a la lista de tuplas del texto etiquetado, obteniendo un árbol de estructuras del cual se extraen los nombres de las entidades.

La técnica básica que hemos usado para la detección de los nombres de las entidades es la estructuración parcial de las frases gramaticales (chunking), que identifica y etiqueta secuencias multi-token como se ilustra en la figura 5. Aunque no se realiza un análisis gramatical completo, sí es posible aislar y etiquetar estructuras de texto complejas que contienen los nombres de las entidades que se buscan[3].

Extracción de entidades

El analizador devuelve un árbol de estructuras. El siguiente paso consistirá en recorrer el árbol e identificar los nombres de entidades reconocidas. En nuestro estudio nos hemos centrado en nombres de instituciones, de hombres, de mujeres y de lugares. La figura 5 muestra el árbol que se obtiene y, finalmente, la lista de 2-tuplas indicando el nombre de la entidad y el tipo de la entidad.

Extracción de relaciones

Identificadas las entidades de interés, nuestro objetivo es determinar las relaciones que aparecen entre ellas. La figura 6 muestra el texto original y las relaciones que han podido obtenerse de ese texto. Este problema no es trivial, ya que, además de identificar las entidades y las relaciones entre ellas, es necesario resolver problemas de coo-referencia y de anáforas. Después de la segmentación y etiquetado del texto, habría que aplicar dos conjuntos de reglas de reconocimiento. Un conjunto de reglas para determinar las entidades y otro para determinar las relaciones entre ellas.

Observemos cómo habría que resolver las anáforas “su mujer” y “mis padres” en la frase “...en España, hijo legítimo de Diego de Chaves y de Magdalena de Velasco, su mujer, mis padres,...”. Las palabras “su mujer” hace referencia a que Diego de Chaves es el marido o esposo de Magdalena de Velasco y que Magdalena de Velasco era la esposa de Diego de Chaves. De igual forma las palabras “mis padres” hace referencia a que Diego de Chaves y Magdalena de Velasco eran los padres del sujeto principal de la oración Eugenio de

[3] Una descripción de esta técnica puede verse en BIRD, S. (…), cap. 7.

Chaves Calizares. Este problema, aunque no se ha resuelto en su totalidad, si hemos llegado a desarrollar varios algoritmos válidos en casos de algunos tipos de anáforas. La información extraída se muestra en tripletas de la forma (sujeto, predicado, objeto) y de esta forma puede representarse en una ontología mediante el lenguaje owl.

Razonadores

La figura 6 además de presentar la información extraída que explícitamente se dice en el texto, muestra información adicional que no se declara en el texto. Por ejemplo, aparece que "Magdalena_de_Velasco per:conyuge per:Diego_de_Chaves". Este tipo de afirmaciones pueden obtenerse tras aplicar a la primera información extraída del reconocimiento un conjunto de reglas que pueden estar explícitamente declaradas en el modelo conceptual de la ontología. Existen lenguajes como swrl o CLIPS, que lo hacen posible. La figura 7 muestra un fragmento de código escrito en lenguaje CLIPS para extender las relaciones familiares identificadas.

Anotación

La siguiente etapa en la arquitectura propuesta en la figura 1 es la anotación de entidades en el documento. El objetivo de esta tarea es modificar el texto del documento, dejando constancia explícita de la parte del texto en la que se ha identificado una entidad. Puesto que la representación del texto de los documentos se ha realizado en xml, se han añadido nuevos elementos al modelo del documento de forma que admita la inclusión de estas nuevas estructuras. La figura 8 presenta el texto de nuestro ejemplo donde se han reconocido y anotado nombres de entidades.

Síntesis

Denominamos estructuras de síntesis a aquellas estructuras de datos organizadas que son obtenidas de las fuentes documentales de forma automática y que facilitan al investigador la elaboración del trabajo final. Ejemplos de esas estructuras podrían ser los nombres de las entidades y de las relaciones reconocidas. Denominamos documentos de síntesis a aquellos documentos que podemos obtener de forma automática y que incorporan estructuras de síntesis cuya misión es facilitar al investigador la elaboración de un trabajo final. La figura 9 muestra las tareas encargadas de realizar la síntesis de estos elementos. A partir de los corpus anotados IOB pueden extraerse las entidades reconoci-

das en las tareas anteriores. Por otra parte, los documentos del corpus anotado xml han sido sometidos a un conjunto de transformaciones xslt, obteniendo de ellos un documento latex y posteriormente un documento pdf que integra los documentos originales, su localización, las estructuras de síntesis que de ellos han sido obtenidas, y un conjunto de índices (onomástico, toponímico, de instituciones, etc.)[4].

Evaluación del sistema

El sistema ha sido sometido a un conjunto de pruebas para validar su comportamiento.

Segmentación, etiquetado y reconocimiento

La identificación de componentes léxicos se ha basado en expresiones regulares y el sistema se ha comportado eficiente logrando un 100 % de precisión. Para el etiquetado de los componentes léxicos hemos utilizado el conjunto de etiquetas propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. El método empleado consiste en la aplicación de un conjunto de reglas de etiquetado basado en expresiones regulares. Aunque han surgido algunos problemas de desambiguación, el sistema se ha comportado con una precisión del 92 %. Una mayor precisión podría obtenerse mediante el empleo de un corpus de documentos etiquetados manualmente y utilizando etiquetadores estadísticos. No obstante, consideramos satisfactorios los resultados obtenidos. Una de las ventajas del método propuesto es independizar el código de la aplicación del fichero que contiene las reglas de etiquetado, esto permite introducir mejoras en el comportamiento del sistema, modificando sólo este fichero, sin alterar el código de la aplicación.

Para evaluar el reconocimiento de entidades, se ha creado un corpus IOB, donde de forma manual, se han identificado todas las entidades del corpus y se ha estudiado diversos parámetros de evaluación.

Se ha utilizado una métrica que incluye varias medidas de evaluación. La figura 10 muestra una representación de los diferentes casos que pueden presentarse al reconocer una entidad:

[4] Un ejemplo de este documento para un conjunto reducido de documentos puede encontrarse en http://www.uco.es/grupos/aaf/projects/pln/documentos/tmp/principal.pdf.

1. TP: (Cierto positivo): Indica que el término ha sido reconocido como un tipo de entidad y es correcto.
2. FP: (Falso positivo): Indica que el termino ha sido reconocido como un tipo de entidad pero es falso.
3. FN: (Falso negativo): Indica que el término no ha sido reconocido, pero en realidad sí debería haber sido reconocido.
4. TN: (Cierto negativo): El término no ha sido reconocido, lo que es cierto ya que no corresponde a una entidad.

- Exactitud (Accuracy): Mide el porcentaje de entradas en el conjunto de prueba que son correctamente reconocidas.
- Precisión: indica cuántos de los elementos que hemos identificado como entidades, realmente lo son: Precision=TP / (TP + FP) es decir el porcentaje de entidades correctamente reconocidas frente al total de las reconocidas. Un sistema será muy preciso si da muy pocas recuperaciones falsas.
- Recubrimiento (Recall): indica cuantos de las entidades que hemos identificado, realmente lo son: Recall=TP / (TP + FN). Es el porcentaje de entidades correctamente reconocidas frente al total de entidades que deberían haberse recocido de ser un sistema perfecto. Un sistema con un Recall=1 sería aquel que no dejara ninguna entidad sin reconocer.
- F-Measure: que combina la precisión y el recubrimiento para dar una única puntuación, se define como la media armónica de la precisión y el recubrimiento: FMeasure=(2 × Precision × Recall) / (Precision + Recall)

Los resultados obtenidos con la gramática utilizada para el reconocimiento de entidades en un conjunto de documentos ha sido el siguiente:

<ChunkScoring of 84 chunks>
ChunkParse score:
IOB Accuracy: 94.1%
Precision: 85.1%
Recall: 67.9%
F-Measure: 75.5%
<ChunkScoring of 84 chunks>

El sistema se comporta de forma eficiente con un bajo número de errores. También aquí hemos independizado el código de la aplicación de las reglas de reconocimiento, con lo que podemos mejorar el rendimiento del sistema modificando el archivo que contiene las reglas gramaticales sin alterar el código de la aplicación.

La representación de los documentos en xml, ha permitido poder anotar explícitamente los nombres de entidades reconocidos mediante elementos xml. Esto favorece, a su vez, poder aplicar a estos documentos transformaciones xslt obteniendo una representación en html o latex. Esta última representación hace posible obtener un documento pdf que integra todos los documentos, incluyendo tablas de estructuras de síntesis y un conjunto de índices onomástico, toponímico y de instituciones de gran utilidad para el investigador.

Conclusiones y futuros trabajos

Hemos desarrollado un sistema que consideramos eficiente para el tratamiento de fuentes documentales con las siguientes características:

La entrada al sistema es un conjunto de documentos escritos en lenguaje natural no estructurado y en formato texto ASCII.

La salida al sistema es un documento pdf en el que se recogen todos los documentos, las tablas de identificación de nombres de entidades y un conjunto de índices relacionados con dichas entidades. Esto hace que el investigador pueda localizar de forma eficiente las partes del texto donde se ubican dichas entidades.

La eficiencia del sistema es alta. Mejoras en los ficheros que contienen las reglas de etiquetado y las reglas de reconocimiento pueden incrementar esta eficiencia.

El usuario no tiene que conocer tecnologías informáticas para poder hacer uso del sistema.

El sistema admite varias mejoras, entre otras destacamos:

Ampliar y perfeccionar el fichero de reglas de etiquetado.

Ampliar y mejorar el fichero de reglas de reconocimiento.

Creación de un corpus documental de documentos etiquetados correctamente de forma que se mejore la eficiencia del etiquetado y se disminuya el problema de desambiguación.

Incorporar un sistema de preguntas y respuestas basado en lenguaje natural que facilite la consulta de la información en el sistema.

Creación de un corpus documental de documentos con un análisis sintáctico parcial correcto de forma que se mejore la eficiencia del reconocimiento.

Bibliografía

[1] Chinchor, N. , Overview of muc-7. In Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998.

[2] A. Toral, A., Dramneri: A free knowledge based tool to Named Entity Recognition, In Proceedings of the 1st Free Software Technologies Conference. A Coruña (Spain). pp. 27-32. July 2005.

[3] Feldman, R. y Sanger., The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge university press, 2007.

[4] Bird, S., Klein, E. and Loper, E., Natural Language Process-ing with Python, O'Reilly Media, 2009.

URL

[5] http://www.nltk.org/Home
[6] http://codespeak.net/lxml/
[7] http://nltk.googlecode.com/svn/trunk/doc/howto/corpus.html
[8] http://www.uco.es/grupos/aaf/projects/pln/documentos.html#normalized
[9] http://www.ilc.cnr.it/EAGLES96/intro.html
[10] http://www.uco.es/grupos/aaf/projects/pln/rules/etiquetas-eagles-1.0.etq
[11] http://www.uco.es/grupos/aaf/projects/pln/rules/gramatica-eagles-entidades.1.1.gr
[12] http://www.uco.es/grupos/aaf/projects/pln/documentos/-tmp/principal.

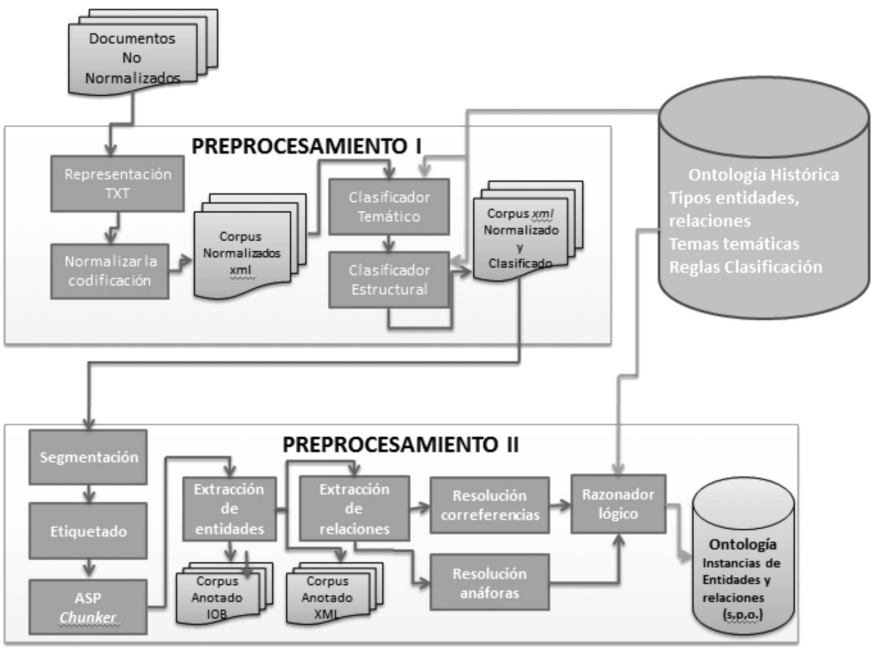

Figura 1 Arquitectura del modelo de la Aplicación

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<documento>
    <persona>Chavez Cañizares, Eugenio de</persona>
    <ciudad>Toledo Yepes</ciudad>
    <lugar>Filipinas Manila</lugar>
    <ayo>1604</ayo>
    <signatura>AGI, Contratación 944B</signatura>
    <tipoDocumento>T</tipoDocumento>
    <texto>
        <p realThemaCat="" strutctCat="" realStrutctCat="" id="0" themaCat="">En el nombre de
Dios amén. Sepan cuantos esta carta de testamento y última voluntad y pstrimera voluntad vieren
como yo, Eugenio de Chaves Calizares, natural de la villa de Yepes, en el Reino de Toledo, en
España, hijo legítmo de Diego de Chaves y de Magdalena de Velasco, su mujer, mis padres,
vecinos que fueron de la dicha villa, ya difuntos.</p>
```

Figura 2 Representación de un párrafo del documento, previo a la clasificación automática

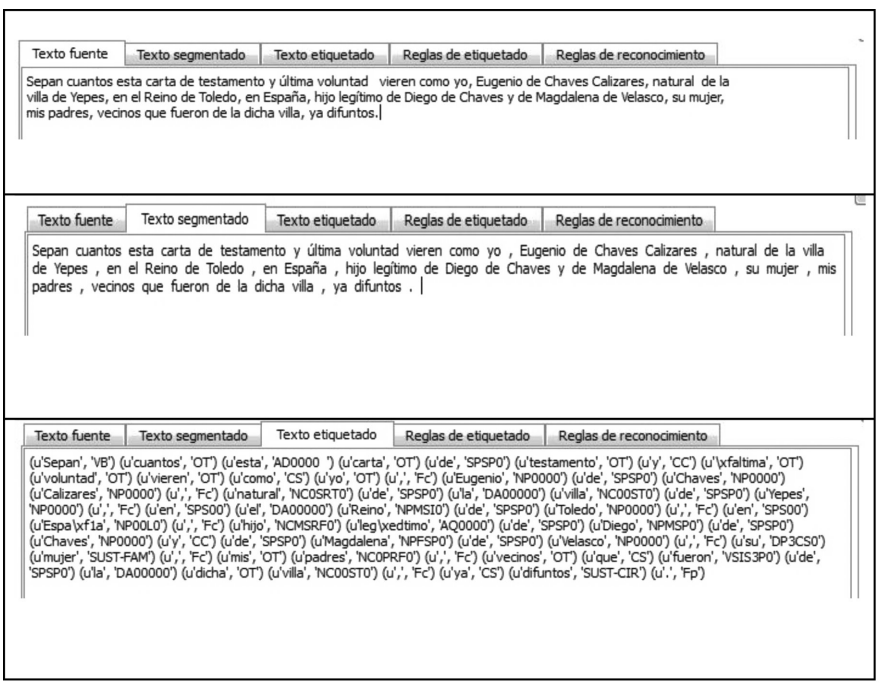

Figura 3 Segmentación y etiquetado de un texto.

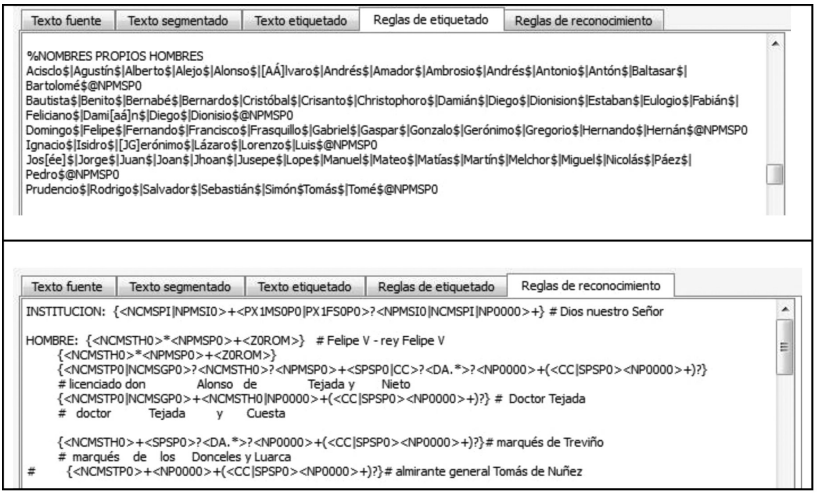

Figura 4 Fragmentos de reglas de etiquetado y reglas de reconocimiento.

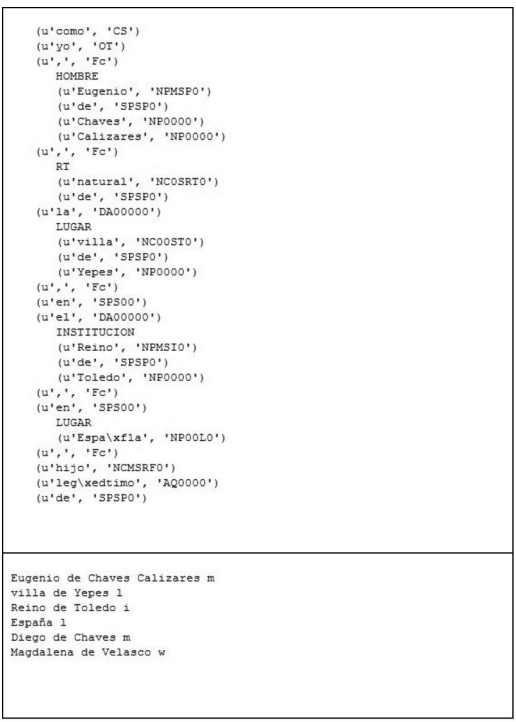

Figura 5 Fragmento del árbol de reconocimiento obtenido y lista de tuplas de entidades obtenidas.

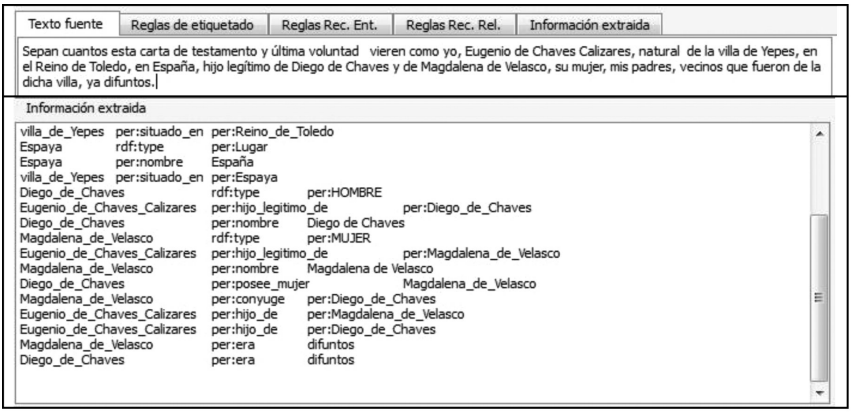

Figura 6 Extracción automática de entidades y relaciones mediante reglas de reconocimiento y la aplicación posterior de un razonador lógico.

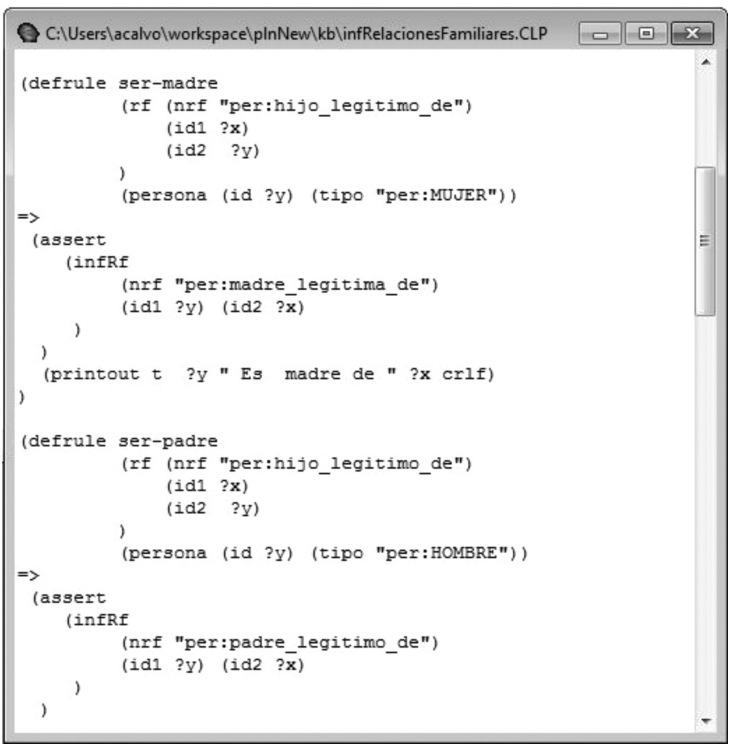

Figura 7 Reglas de inferencia en lenguaje CLIPS que aplica el razonador a la información extraída automáticamente (entidades y relaciones) extendiendo las relaciones o detectando incoherencias.

```
<p realThemaCat="" strutctCat="" realStrutctCat="" id="0" themaCat="">En el nombre de Dios
amén. Sepan cuantos esta carta de testamento y última voluntad y pstrimera voluntad vieren como
yo, Eugenio de Chaves Calizares, <rt>natural de</rt> la <l>villa de Yepes</l>, en el Reino de
Toledo, en <l>España</l>, hijo legítmo de <m>Diego de Chaves</m> y de <w>Magdalena de
Velasco</w>, su mujer, mis padres, vecinos que fueron de la dicha villa, ya difuntos.</p>
```

Figura 8 Anotación explícita de los nombres de los elementos.

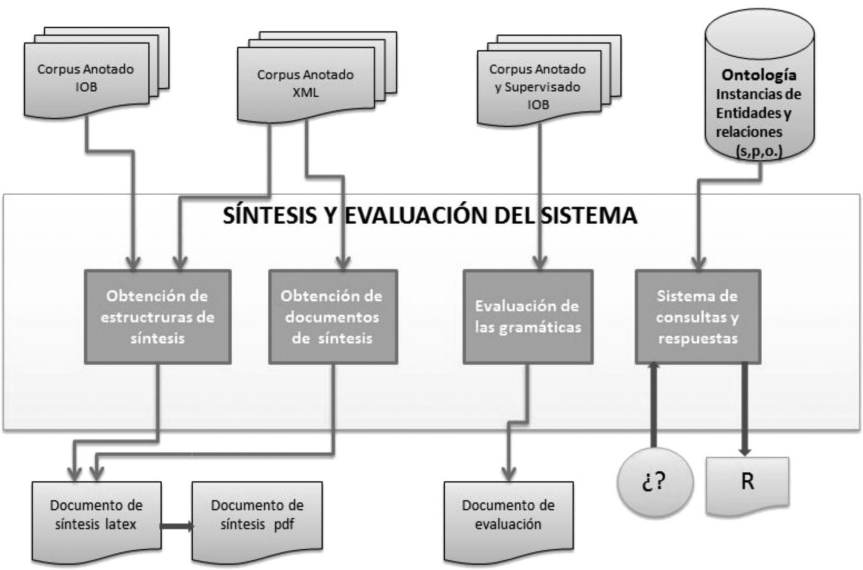

Figura 9 tareas para la síntesis y evaluación del sistema.

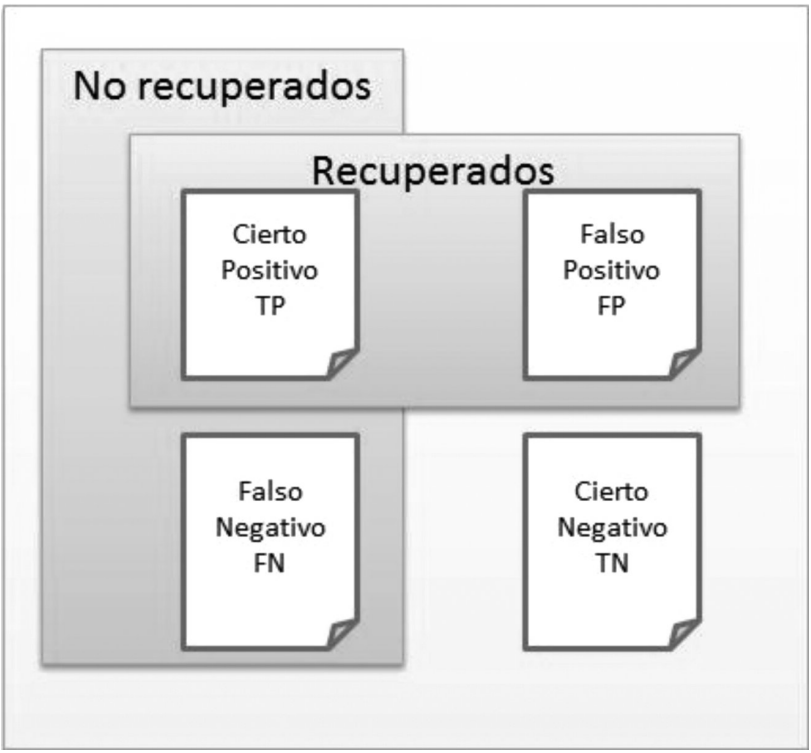

Figura 10 Parámetros para definir una métrica de evaluación.