Generación semiautomática de una ontología para una red de ONG

Natalia Chiaro
nchiaro@internet.com.uy

Pablo Damonte
pdamonte@montevideo.com.uy

Diego Garat
dgarat@fing.edu.uy

Pablo Accuosto
accuosto@chasque.net

Facultad de Ingeniería
Universidad de la República
Uruguay

Resumen

Para que el conocimiento almacenado en la *Web* pueda ser efectivamente recuperado y utilizado en forma automática, es necesario enriquecer las páginas con *metadatos* que permitan definir conceptos y relaciones en un dominio específico. Este dominio se representa mediante una *ontología*, mediante la cual se definen las entidades relevantes y las relaciones que las vinculan.

Este trabajo busca encontrar e implementar técnicas eficaces que permitan, con la menor intervención del usuario, generar una ontología a partir de documentos publicados en un sitio *Web* de una red de organizaciones no gubernamentales. La solución propuesta parte de una ontología mínima, construida manualmente, que se completa con entidades nombradas y relaciones identificadas automáticamente.

La evaluación del grado de reconocimiento y precisión en la recuperación de entidades y relaciones de la herramienta implementada permite estimar que las técnicas propuestas pueden constituir un aporte relevante para la generación semiautomática de ontologías.

Palabras claves: Web semántica, generación de ontologías, extracción de información.

1 Introducción

La «*Web Semántica*», propuesta inicialmente por Tim Berners-Lee [1], es actualmente una iniciativa del *World Wide Web Consortium* (W3C) [2]. Concebida como la siguiente etapa en el desarrollo de la *Web*, parte de la base que ésta sólo puede alcanzar su pleno potencial si el conocimiento contenido en ella, que actualmente es «entendible» —en general— únicamente por humanos, puede ser compartido y procesado por herramientas automáticas. Para lograr este fin, es necesario enriquecer y estructurar la información disponible actualmente en la *Web*, lo cual implica acordar formatos para la representación de este conocimiento y mecanismos que posibiliten su utilización eficaz por parte de aplicaciones desarrolladas de forma independiente.

Para que pueda existir un procesamiento semántico de los documentos en un escenario *Web*, idea fundamental de la *Web Semántica*, es necesario tener en cuenta tres elementos fundamentales. El primero de ellos es contar con agentes de software con la capacidad de procesar automáticamente documentos, realizando análisis semántico de su contenido. El segundo es la existencia de un modelo conceptual que describa los rasgos característicos de un dominio dado, entidades y relaciones entre ellas, mediante una *ontología*. El último consiste en la existencia de *metadatos*: información que se asocia a los documentos para describir su contenido semántico.

La calidad y la correcta utilización de todos estos elementos posibilitarán la mejora en la localización y procesamiento de los datos de las páginas pertenecientes a un dominio específico mediante el uso de herramientas informáticas.

El presente trabajo busca generar de forma semiautomática una ontología, utilizando para ello una especificación estándar de Web Semántica. Con tal fin, se analizaron y evaluaron distintas técnicas para la extracción y categorización semiautomática de información.

Se trabaja sobre un dominio particular, ***Choike***[1], un portal destinado a mejorar la visibilidad de los contenidos producidos por organizaciones no gubernamentales (ONG); este portal contiene informes periodísticos sobre temas económicos, políticos y sociales. Luego, se plantea el objetivo concreto de construir un prototipo que a partir de un conjunto de estos informes genera la ontología de Choike.

El presente trabajo se estructura en siete secciones. En la sección 2, se presenta una breve introducción a la problemática de la Web semántica. En las secciones 3, 4 y 5, se describe el problema de extracción en el dominio particular elegido, la solución propuesta y la descripción de su arquitectura. El prototipo implementado se evalúa en la sección 6. Finalmente, en la sección 7, se presentan las conclusiones y los trabajos a futuro.

2 La Web semántica

Para que la *Web Semántica* se vuelva efectiva, el intercambio de metadatos debe ser realizado teniendo en cuenta los aspectos relevantes a la interoperabilidad semántica, sintáctica y estructural. La comprensión de los descriptores, provenientes de distintas fuentes, y sus relaciones es posible gracias a la interoperabilidad semántica, lograda por el uso de vocabulario específico, ontologías y estándares para metadatos.

Los metadatos nos permiten tener información extra de los datos, y su diseño se encuentra por lo general influenciado por una ontología. La figura 1 muestra la relación entre las ontologías, metadatos y los propios datos.

Figura 1 – Relaciones

Por otra parte, existe una gran diversidad de lenguajes de descripción de ontologías, cada uno con sus características específicas, permitiendo diferentes niveles de expresividad a la hora de describir los elementos que la componen: conceptos, relaciones, funciones, axiomas e instancias.

Algunos de los lenguajes existentes brindan un grado de expresividad limitado, al no permitir asociar semánticamente sus etiquetas, como es el caso de *RDF-Schema* [3]. El lenguaje OIL[5] avanza un poco más, y permite definir un vocabulario al cual se le asocia un significado que es entendido a nivel de máquina, aunque carece de mecanismos para expresar negaciones o disyunciones [4].

En un nivel más avanzado, se encuentran aquellos lenguajes que proveen de mecanismos para la declaración de propiedades referidas a los recursos, relaciones y clases, los que poseen un modelo de herencia bien definido o los que permiten establecer relaciones más complejas entre las entidades —mecanismos para limitar las propiedades de las clases respecto al número y al tipo— o

[1] http://www.choike.org/

inferencias para determinar la clase del objeto a partir de sus propiedades. Dentro de estos últimos se puede mencionar al estándar *OWL* [6].

Existen lenguajes, por ejemplo *CKML* [7], que además permiten la representación de conceptos organizados en taxonomías, relaciones o axiomas de lógica de primer orden. Otros, como *CycL* [8], van un poco más allá, y permiten expresar conceptos de una lógica de mayor orden.

La iniciativa *DAML+OIL* [9] apunta a proporcionar un lenguaje y un conjunto de herramientas que habiliten la transformación de la *Web:* de una plataforma que presenta información a una plataforma que entienda y razone, que incluso pueda soportar una semántica declarativa, en la que el significado de las expresiones en una representación puede ser entendido sin necesidad de recurrir a un intérprete para su manipulación.

Si bien existe una gran heterogeneidad y diversidad de opciones a la hora de elegir un lenguaje, no todos son adecuados o directamente aplicables en cualquier contexto, habiendo lenguajes específicos para un área de aplicación y otros de propósito más general.

Un aspecto importante a considerar es el hecho de que muchos lenguajes se implementan en base a otros (figura 2), lo que garantiza cierta compatibilidad entre los distintos lenguajes, al menos en una dirección.

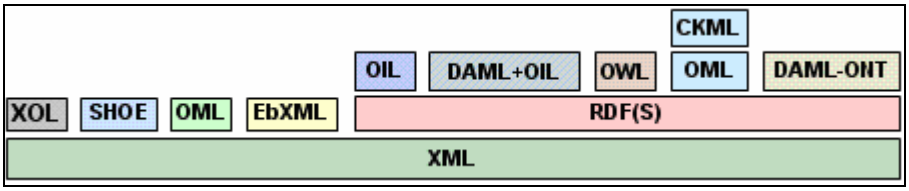

Figura 2 - Jerarquía de lenguajes en la Web Semántica

En particular, y luego de haber analizado y estudiado los distintos lenguajes y herramientas que permiten definir y utilizar las ontologías, se decide utilizar al lenguaje *OWL* [6] para llevar a cabo el trabajo planteado. Durante la evaluación de los lenguajes antes mencionados, se hace hincapié en la variedad de posibilidades que estos brindan, como ser la de representar conceptos y relaciones, su facilidad de uso, etc. También se considera el nivel soporte que a estos les dan a las distintas herramientas que actualmente se encuentran disponibles en el mercado, su grado de estandarización y el nivel de aceptación que reciben.

3 Ontología de Choike

El problema a resolver es la generación automática de una ontología representativa del dominio de Choike, utilizando la información publicada en su sitio Web. Este sitio tiene como objetivo la difusión de material elaborado por las ONG de todo el sur, así como de material que pueda ser de interés de éstas. En particular, en Choike se publican informes periodísticos sobre temas políticos, económicos y sociales desde el punto de vista de la sociedad civil. Considerando el volumen de información que presentan, es en base a estos documentos que se decide construir la ontología.

Según la temática de los informes, y luego de su análisis, se determina cuales son las principales entidades y relaciones que deben conformar la ontología de Choike. Entre las entidades se destacan las personas, cargos, países, organizaciones, documentos y eventos; mientras que entre las relaciones consideradas se encuentran la vinculación de una persona a una organización, la de una organización a un país, que una persona esté ejerciendo o haya ejercido algún cargo, etc. En la figura 3, se encuentran las distintas relaciones consideradas relevantes junto con una breve descripción de éstas.

Abreviación	Asocia una organización, evento o documento con su correspondiente sigla o acrónimo. Ejemplo: "En el Fondo Monetario Internacional (*FMI*) se estudia la….." Relación: *Abreviación* (Fondo Monetario Internacional, *FMI*)
Vinculado_a	Vincula una persona a un documento, organización o evento. Ejemplo: "El presidente del Comité Pacificación, Juan Pérez, fue quien…." Relación: *Vinculado_a* (Juan Pérez, Comité Pacificación)
Pertenece_a	Vincula una organización, evento o documento a un país. Ejemplo: "Universidad de la República (Uruguay)…." Relación: *Pertenece_a* (Universidad de la República, Uruguay)
Persona_Cargo	Asocia una persona a un determinado cargo. Ejemplo1: "El presidente del Comité Pacificación, Juan Pérez, fue quien…." Relación: *Persona_Cargo* (Juan Pérez, presidente) Ejemplo2: "El ex-ministro de cultura Pedro Fernández…" Relación: *Persona_Cargo* (Pedro Fernández, ex-ministro)
Ciudad_de	Vincula una ciudad con el país al que pertenece. Ejemplo: "…la cual fue dictada en Kyoto, Japón." Relación: *Ciudad_de* (Kyoto, Japón)

Figura 3 – Descripción de las relaciones de la ontología

Además, se decide reflejar en la ontología el vínculo entre cada entidad o relación y los informes en los cuales efectivamente ocurren. Esta información se agrega con el fin de ser explotada en una siguiente etapa, por ejemplo, en la búsqueda de informes, en la recomendación al lector por similitud entre informes, etc. En la figura 4 se puede apreciar los distintos tipos de entidades y relaciones contempladas.

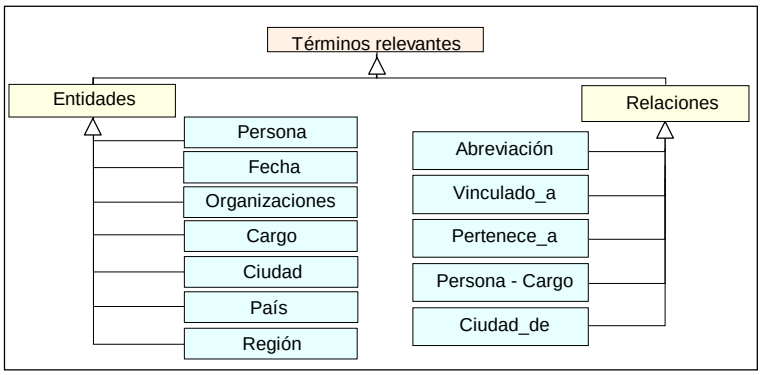

Figura 4 - Entidades y relaciones

Por otra parte, los informes se encuentran clasificados en cinco grandes categorías —«la gente», «la sociedad», «el ambiente», «la comunicación» y «la globalización»— divididas a su vez en dieciocho subcategorías —«afrodescendientes», «biodiversidad», «comercio e integración regional», etc.—. Esta categorización temática en «dos niveles» de los informes es también incorporada a la ontología.

En consecuencia, se concibe una *ontología inicial* u *ontología base* con las entidades y relaciones relevantes del dominio (figura 5), agregándose algunas instancias de organizaciones y países obtenidos de diccionarios.

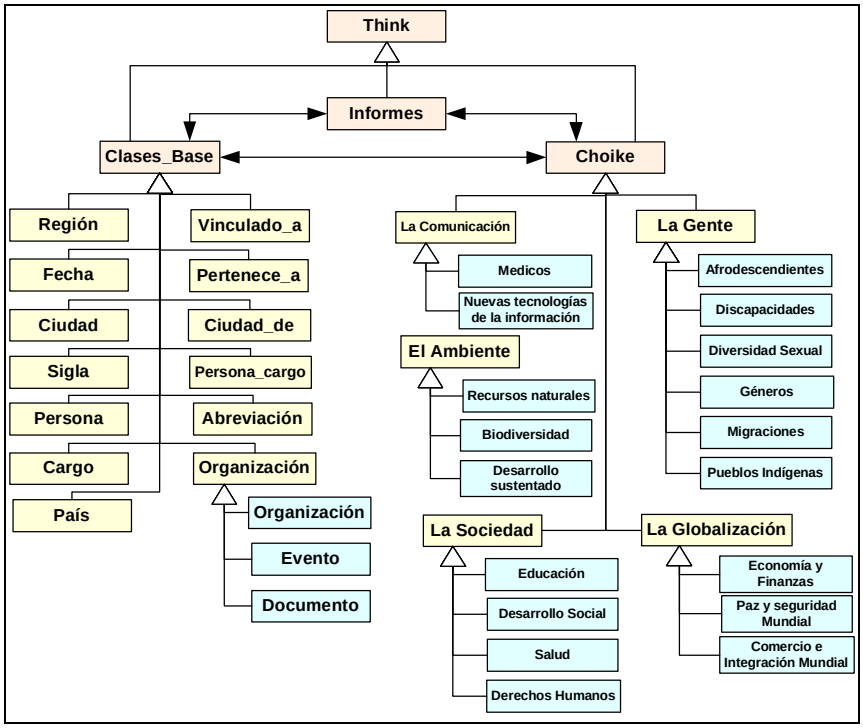

Figura 5 - Ontología base de Choike

Esta ontología base es una de las entradas de la herramienta de extracción, la cual se encarga de aumentarla con las instancias de entidades y relaciones que detecta en los informes de las páginas de Choike.

4 OntoChoike

Como se menciona en la sección anterior, la solución planteada, denominada *OntoChoike*, toma como entrada un conjunto de páginas Web que contienen informes de *Choike* y una ontología base. OntoChoike extrae las entidades y relaciones de los informes, incorporándolas a la ontología inicial (figura 6).

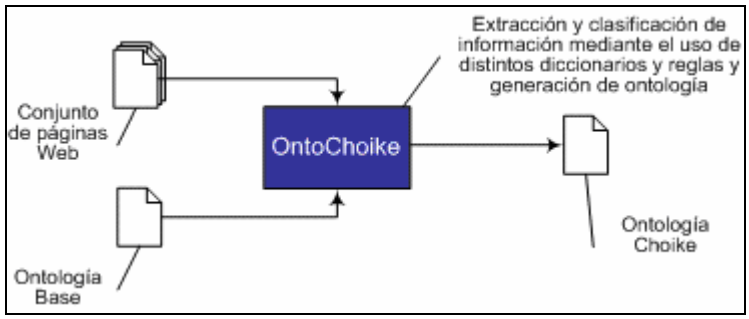

Figura 6 - Solución Propuesta

En principio, podría considerarse suficiente el tener un gran diccionario que contenga los nombres de entidades (nombres propios, países, etc.) y reconocer las entidades realizando su búsqueda en el texto. Sin embargo, esta solución no es satisfactoria: el conjunto de entidades a reconocer no se puede considerar «cerrado». Por ejemplo, los nombres de ciudades o países pueden considerarse como algo invariante (partiendo de la base que los países o ciudades no se crean o cambian de nombre en forma muy habitual), pero los nombres de personas u organizaciones distan

mucho de serlo. Se hace necesario, entonces, el reconocimiento de nuevas entidades, esto es, entidades de los cuales no se tiene información previa.

La solución propuesta, a partir de diccionarios de «palabras disparadoras» y reglas, reconoce un conjunto abierto de nombres de entidades y relaciones. Los diccionarios permiten determinar si una palabra es candidata a formar parte de una entidad. Así, por ejemplo, la palabra «asociación» es marcada como un posible comienzo de una organización. En cambio, en la siguiente etapa, las reglas hacen distintas validaciones que posibilitan tener un mayor grado de exactitud en los datos reconocidos, ayudando a reconocer información adicional o filtrando, por ejemplo, fechas inválidas. Finalmente, se detectan relaciones entre las entidades previamente clasificadas

La tarea de extracción se divide, entonces, en tres etapas diferenciadas: detección de candidatos a entidades, filtrado y construcción de los candidatos y detección de relaciones. A continuación se detallan cada una de estas etapas.

4.1 Detección de términos relevantes

Esta etapa tiene como objetivo detectar en los informes de Choike las palabras que son potenciales entidades de la ontología, utilizando para esto diccionarios de «palabras disparadoras». Estas palabras indican la posible presencia de una entidad a ser reconocida, y dependen del tipo de entidad considerada (país, organización, etc.); por esto, se agrega un diccionario de «palabras disparadoras» por cada clase de entidad a reconocer.

El proceso consiste, entonces, en detectar y extraer de los textos «palabras disparadoras», agregando, para cada una de ellas, información referente al tipo potencial de entidad de la que pueden ser parte, en qué posición fueron encontradas y la ventana de palabras que las rodean.

La ventana de palabras permite formar el nombre completo de las entidades en un proceso posterior. Su tamaño varía según el tipo estimado de la entidad: por ejemplo, en el caso que sea una organización, la ventana es de treinta palabras, mientras que en el caso de un nombre de persona se utilizan diez.

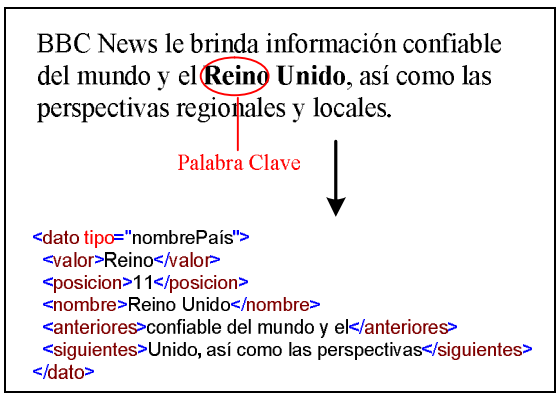

Figura 7 - Ejemplo de Reconocimiento

Adicionalmente, para los casos de ciudades y países, se agregan los distintos nombres a los que la palabra detectada puede ser parte. Por ejemplo, al encontrarse la palabra «Reino», dado que esta es una palabra clave asociada al nombre de país «Reino Unido», se le asigna el tipo de entidad candidato *País* y, además de agregarse las cinco palabras anteriores y cinco próximas, se agrega información referente a que «Reino Unido» es el posible nombre del país que se está reconociendo (ver figura 7).

Es importante remarcar además que los diccionarios de ciudades y países, no sólo contienen la palabra clave que es usada para reconocer las entidades, sino que además contiene su nombre completo; a partir del diccionario, se puede saber que *Los Ángeles* tiene como palabra clave a *Ángeles* y además pertenece a *Estados Unidos*.

Cuando la palabra considera pertenece a más de un diccionario, surge un problema de ambigüedad. Estos casos se resuelven otorgando un orden de precedencia a los distintos tipos de instancias. Por ejemplo, la palabra *Argentina* puede ser tanto el nombre de un país o el de una persona. OntoChoike le da preferencia al nombre del país, con lo que *Argentina* siempre denota a un país. Esta opción se justifica por el tipo de entidades que ocurren en el dominio específico de Choike.

4.2 Construcción de entidades

Luego de extraer las palabras clave junto a su contexto dentro del documento, se aplica un proceso que tienen un doble objetivo: por una parte, se filtran aquellas palabras que, a pesar de ser disparadoras, no forman parte de una entidad; por otra, se construye la totalidad de las entidades. Esta tarea se realiza aplicando un nuevo conjunto de reglas a la información extraída en el paso previo.

Un ejemplo de regla de inferencia aplicada en esta instancia del proceso es la utilizada para reconocer nombres de personas, la cual consiste en asumir que un nombre de pila, seguido de letras capitales, palabras comenzadas en mayúscula o palabras «aglutinantes» como ser «de», «del», etc. son parte del nombre de una persona. De esta forma se puede reconocer que *Jorge del Campo* es el nombre de una persona, dado que *Jorge* pertenece al diccionario de nombres, *del* es una palabra contemplada en los nombres de personas y ésta, a su vez, está seguida de otra palabra comenzada en mayúscula.

Figura 8 - Ejemplo Entidad

Por ejemplo, en la etapa previa, dado el texto de entrada «Juan Pablo Pérez realizó...», se reconocen las palabras claves «Juan» y «Pablo», y ambas son clasificadas, de forma independiente, como potenciales constituyentes de un «nombre». En esta etapa, las reglas detectan la ocurrencia de dos «nombres» consecutivos, y los agrupa en una única entidad. Además, se continúa la entidad abarcando a la siguiente palabra del contexto, «Pérez», por comenzar en mayúscula, a pesar de no pertenecer al diccionario de nombres.

El proceso que se realiza para reconocer nombres de organizaciones, eventos y documentos es similar al de personas, utilizando, además de un diccionario distinto, un conjunto diferente de partículas «aglutinantes».

4.3 Extracción de relaciones

A partir de las entidades reconocidas, se establecen relaciones de distinto tipo entre ellas. Se buscan reglas que reflejen las relaciones y que requieran del menor costo posible de análisis.

Este reconocimiento varía de acuerdo al tipo de relación que se está reconociendo y, en gran medida, a cómo estas relaciones ocurren en los informes de *Choike*.

El método utilizado para detectar los distintos tipos de relaciones consiste en el uso de diversas reglas. Por ejemplo, la relación *Vinculado_a* es reconocida cuando se encuentra una instancia de una organización, evento o documento y una instancia de una persona a una distancia menor a 20 palabras.

Como ejemplo, en el texto de la figura 9, se reconocen a la persona «Carlos Fernández» y a la organización «Universidad de Montevideo», y como su distancia es menor a 20, se establece entre ellas la relación *Vinculado_a*.

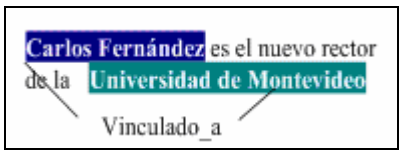

Figura 9 - Ejemplo Relación Vinculado_a

Figura 10 - Ejemplo Relación

En otros casos, como la relación *Abreviación*, se apuesta fuertemente a las reglas que reconocen esta relación cuando encuentran el nombre de una organización, evento o documento, seguido de una sigla entre paréntesis o de un guión (-) y una sigla. Un ejemplo de esto se da en la figura 10.

5 Implementación

La *ontología inicial*, siguiendo el estándar OWL, se crea utilizando *Protégé* [10]. A esta se le agregan datos adicionales reconocidos y clasificados de las páginas del sitio, utilizando el sistema que se implementa utilizando *Java*. Además, para el manejo de *XML* se utiliza *XPATH* [13], y para *XSLT*, el paquete *JDOM* [12]. Como resultado del proceso se obtiene un archivo *OWL* conteniendo la ontología generada.

La división del problema original en dos subproblemas da lugar a las dos etapas en las que se divide el funcionamiento del motor que conforma la herramienta de extracción. Estas etapas, esquematizadas en la figura 11, son:

- **text2xml**: procesa las páginas *web* con los informes, extrae la información reconocida y genera un archivo *XML*.
- **xml2owl**: procesa las entidades reconocidas en la etapa anterior y genera la ontología en formato *OWL*.

A continuación, se describen cada una de las etapas mencionadas.

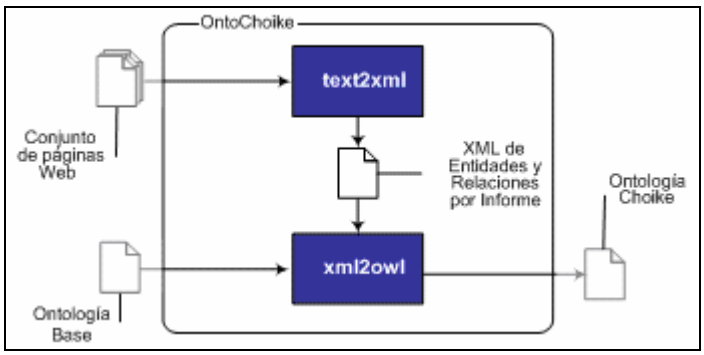

Figura 11 - Arquitectura OntoChoike

5.1 text2xml

El cometido principal de este componente es reconocer y clasificar las entidades y relaciones contenidas en los informes, mediante las reglas definidas en la sección 4. La etapa se encuentra dividida en los procesos de extracción del texto, *tokenización*, reconocimiento de términos relevantes, filtrado y, por último, reconocimiento de relaciones. Un esquema de su arquitectura se puede apreciar en la figura 12.

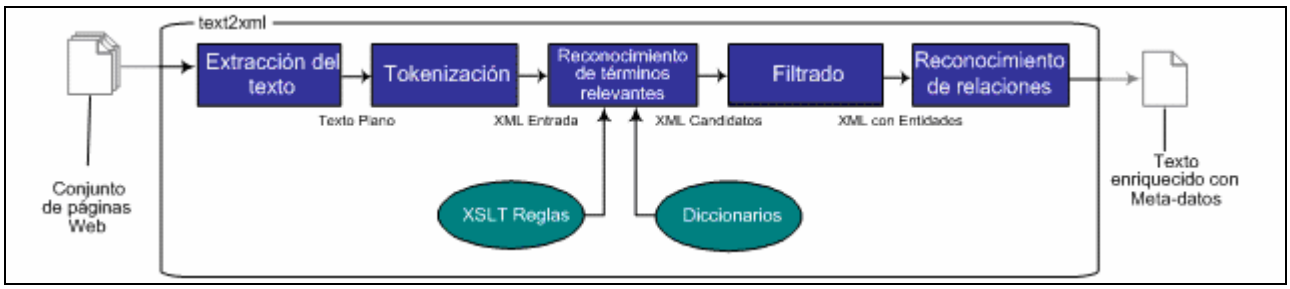

Figura 12 - Arquitectura text2xml

- **Extracción de texto.** – Este proceso extrae el texto del informe contenido en una página Web, omitiendo aquellas partes de la página que no contienen información relevante (cabezal, pie de página, etc.). Este proceso se implementa con el paquete *HTMLParser* [11], el cual brinda soporta para el manejo de páginas *HTML*.
- **Tokenización.** – El propósito de este proceso es la generación de un documento *XML* (mediante el uso de *JDOM*) a partir del texto plano de entrada, el cual contendrá la misma información, pero con una estructura de *XML*.
- **Reconocimiento de términos relevantes.** – Este proceso hace uso de *XSLT* para descartar los datos (nodos) que no aportan información, dejando para su posterior procesamiento sólo aquellos datos de los cuales se puede obtener información relevante. Para realizar este filtrado se aprovecha el soporte que *XSLT* da al uso de expresiones regulares.
- **Filtrado.** – En esta etapa se procesa el documento *XML* obtenido en la etapa previa para terminar de filtrar y reconocer entidades y relaciones, usándose para esto *JDOM* y expresiones regulares. Luego, se finaliza la incorporación de toda la información encontrada para un cierto nombre de persona, fecha, etc.
- **Reconocimiento de relaciones.** – Es en esta etapa donde, a partir de las entidades antes reconocidas y filtradas, se reconocen las relaciones existentes entre ellas.
- **Salida.** – La salida de este proceso es un archivo *XML*, donde se presentan los datos encontrados por página (informe), de qué tipo son y las relaciones que pueden tener con otras entidades.

5.2 xml2owl

Esta etapa tiene como principal objetivo la generación de una ontología en formato *OWL*, a partir de un archivo *XML* de entrada, salida de *text2xml*, y una ontología base. La modificación de la ontología base se hace mediante el uso de *JDOM*, teniéndose en cuenta todos los recaudos necesarios para que el archivo generado sea completamente compatible con *Protégé*.

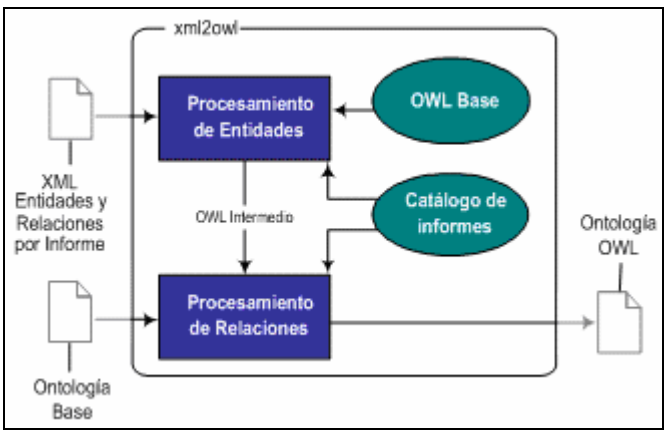

Figura 13 - Arquitectura xml2owl

La etapa se encuentra dividida en dos procesos. El primero de ellos se encarga de la realización del procesamiento de entidades y el segundo efectúa posteriormente el procesamiento de las relaciones encontradas.

6 Resultados

A continuación, se presentan los resultados obtenidos con la solución *OntoChoike*. Cabe señalar que todas las pruebas fueron realizadas en una computadora de escritorio equipada con un procesador AMD XP 2000+, 640 MB de RAM, corriendo el sistema operativo Windows XP Professional SP2 y Java Virtual Machine 1.4.2.

Los resultados obtenidos fueron analizados según las dos etapas que conforman al problema. Para el caso de *xml2owl*, lo importante es el volumen de los datos que contiene la ontología generada, mientras que en lo referente a *text2xml* la evaluación se centró en analizar la calidad y correctitud de los datos extraídos: de obtenerse un tamaño considerable de datos, pero incorrectos, no sólo no ayuda a generar una ontología específica, sino que además puede degenerar la *ontología base* en una incorrecta.

6.1 Forma de evaluación

Para poder realizar la estimación de los errores cometidos al reconocer información, es necesario reconocer y clasificar la información relevante en forma manual. Por tal motivo se tomaron en forma aleatoria, veinte informes de *Choike* para ser usados como *corpus*, con la consiguiente clasificación manual de sus datos relevantes.

Estos informes fueron divididos en dos grupos de diez informes cada uno. Los primeros diez, se utilizaron para afinar el sistema: a los diccionarios y reglas que fueron definidos en función del contenido y estructura global de los informes del sitio se les hicieron las modificaciones necesarias con el objetivo de reconocer la mayor cantidad de información relevante, cometiendo a su vez el menor error posible. Los restantes diez informes fueron utilizados para la realización de la evaluación.

Para estimar los errores cometidos en el reconocimiento, se usaron las medidas de *recuperación* (que determina la cantidad de datos reconocidos) y *precisión* (que mide la calidad de los datos reconocidos), siendo sus fórmulas las siguientes:

$$\textit{recuperación} = VP / (VP + FN) \qquad \textit{precisión} = VP / (VP + FP)$$

Donde

VP = Instancias reconocidas correctamente por la aplicación
FN = Instancias no reconocidas, pero que debían haberse reconocido
FP = Instancias reconocidas erróneamente por la aplicación

Adicionalmente, se usó la medida de *F-measure* (con $\alpha = 0{,}50$) que combina las dos medidas anteriores:

$$\textit{F-measure}(\alpha) = (\alpha \,.\, \textit{precisión}^{-1} + (1-\alpha) \,.\, \textit{recuperación}^{-1})^{-1}$$

6.2 Resultados obtenidos

Como se mencionó en la sección anterior, un subconjunto del *corpus* se utiliza para estimar la *precisión*, *recuperación* y *F-measure* de la etapa. Para estos diez informes, se insumió un tiempo total de procesamiento de 2 minutos 14 segundos. Cabe destacar que no se hizo mayor hincapié en los tiempos en que incurría el proceso en su conjunto, debido a que éste no era un tema planteado como relevante: la extracción y clasificación de información a partir de las páginas no es una tarea a realizar en forma asidua.

En el reconocimiento de entidades se obtienen muy buenos resultados, con una *F-measure* del 84%. (figura 14). En el caso del reconocimiento de relaciones, los resultados son sensiblemente inferiores a los obtenidos en el reconocimiento de entidades, con una *F-measure* del 69%. (figura 15).

Este último resultado se encuentra directamente afectado por el error al reconocer entidades, dado que las relaciones se detectan a partir de las entidades reconocidas en el paso previo. Si se estima el error utilizando como entrada todas las entidades presentes en el texto correctamente etiquetadas, se observa un aumento tanto en la recuperación como en la precisión del resultado, llegando a valores superiores a los obtenidos en la extracción de entidades (columnas marcadas con (*) en la figura 15).

Entidades	FN	VP	FP	R	P	F-m
organización	56	97	24	0,63	0,80	**0,71**
documento	11	42	13	0,79	0,76	**0,78**
evento	5	32	8	0,86	0,80	**0,83**
país	0	81	1	1,00	0,99	**0,99**
ciudad	4	24	1	0,86	0,96	**0,91**
persona	15	8	4	0,35	0,67	**0,46**
fecha	14	22	0	0,61	1,00	**0,76**
región	1	12	3	0,92	0,80	**0,86**
cargo	6	16	0	0,73	1,00	**0,84**
sigla	0	99	3	1,00	0,97	**0,99**
Total	**112**	**433**	**57**	**0,79**	**0,88**	**0,84**

Figura 14 - Resultados del reconocimiento de entidades

Relaciones	FN	VP	FP	R	P	F-m$_{[0.5]}$	FN*	VP*	FP*	R*	P*	F-m$_{[0.5]}$*
abreviación	10	20	4	0,67	0,83	**0,74**	5	25	4	0,83	0,86	**0,85**
pertenece a	24	23	8	0,49	0,74	**0,59**	5	42	8	0,89	0,84	**0,87**
vinculado a	5	4	9	0,44	0,31	**0,36**	2	7	4	0,78	0,64	**070**
ciudad de	5	24	1	0,83	0,96	**0,89**	1	28	0	0,97	1,00	**0,98**
persona cargo	2	3	0	0,60	1,00	**0,75**	2	3	0	0,60	1,00	**0,75**
Total	**46**	**74**	**22**	**0,62**	**0,77**	**0,69**	**15**	**105**	**16**	**0,88**	**0,87**	**0,87**

Figura 15 - Resultados del reconocimiento de relaciones

Considerando la aplicación en su conjunto, los errores cometidos al reconocer entidades y relaciones son:

recuperación: 0,76 *precisión*: 0.87 *F-measure* (0.5): 0,81

En total se reconocieron 41.780 ocurrencias de entidades y relaciones (36.659 entidades y 5.121 relaciones) luego de procesar la totalidad de los informes (1189), en un tiempo total de 2 horas 40 minutos.

7 Conclusiones y trabajo futuro

En este trabajo se presenta una solución incremental al problema generación semiautomática de ontologías para el dominio específico de Choike. Se implementa un prototipo que genera una ontología extrayendo información a partir de un conjunto de informes, encontrándose, entonces, una solución a una tarea compleja mediante la utilización de técnicas de extracción para un escenario *Web*. La herramienta construida presenta muy buenos niveles de recuperación y precisión dentro del dominio de aplicación.

Como trabajo a futuro se plantea la mejora al conjunto de reglas de inferencia para reconocer entidades y relaciones. Esto implica no solo aumentar la precisión de las reglas, sino también su número, con el objetivo de incorporar casos hasta el momento excluidos, mejorando con esto la recuperación global del sistema.

Por otro lado, el usuario tiene poca capacidad de decisión sobre el proceso de extracción en el sistema actual. Algunas posibles opciones tendientes a subsanar este problema son las siguientes: (a) poder seleccionar de las entidades y relaciones ya existentes en los diccionarios cuáles se desea

extraer dinámicamente en el momento de la generación; y (b) brindar la posibilidad de ampliar los diccionarios que utiliza el sistema, ya sea de forma manual o semiautomática, para así poder reconocer nuevas entidades y relaciones.

Finalmente, cabe preguntarse qué niveles de respuesta se obtienen en otros dominios de trabajo disímiles al de Choike. Se plantea, entonces, realizar una evaluación sobre otro conjunto de textos, provenientes de fuentes heterogéneas. Esta evaluación permitirá estimar el grado de adaptabilidad de la herramienta construida.

8 Bibliografía

[1] A roadmap to the Semantic Web (Sept 98) Tim Berners-Lee
http://www.w3.org/DesignIssues/Semantic.html
Fecha de acceso: 11/05/2004

[2] Web Semántica
http://www.w3.org/2001/sw/
Fecha de acceso: 11/05/2004

[3] W3C – Estándar RDF-Schema.
http://www.w3.org/XML/
Fecha de acceso: 29/06/2005

[4] Luke, S.; Spector, L.; Rager, D. Ontology-Based Knowledge Discovery on the World-Wide Web. Proceedings of the Workshop on Internet-based Information Systems, AAAI-96 (Portland, Oregon), 1996

[5] Fensel, D. et al. (2000). OIL in a nutshell. Knowledge Acquisition, Modeling, and Management, Proceedings of the European Knowledge Acquisition Conference (EKAW-2000), R. Dieng et al. (eds.), Lecture Notes in Artificial Intelligence, LNAI, Springer-Verlag, October 2000.

[6] OWL Web Ontology Language Overview.
http://www.w3.org/2001/sw/WebOnt/TR/STAGE-owl-features/
Fecha de acceso: 24/05/2004

[7] R.E. Kent. *Conceptual Knowledge Markup Language: The Central Core*, in: Twelfth Workshop on Knowledge Acquisition, Modeling and Management (1999).

[8] CycL.
http://www.cyc.com/cycl.html
Fecha de acceso: 24/05/2004

[9] McGuinness, Deborah; Fikes, Richard; Stein, Lynn; Hendler, James. DAML+OIL: An Ontology Language for the Semantic Web ". In IEEE Intelligent Systems, Vol. 17, No. 5, pages 72-80, September/October 2002

[10] Protégé
http://protege.semanticweb.org/
Fecha de acceso: 11/10/2004

[11] HTMLParser
http://www.htmlparser.org/
Fecha de acceso: 13/03/2005

[12] JDOM
http://www.jdom.org/
Fecha de acceso: 13/03/2005

[13] XPATH
http://www.w3.org/TR/xpath
Fecha de acceso: 1/04/2005