Construcción automática de diccionarios de patrones de extracción de información

Neus Català, Núria Castell
Dept. Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1-3, 08034 Barcelona
Edificio C6, Campus Nord
Tel: (+3) 4016994; Fax: (+3) 4017014
ncatala@lsi.upc.es, castell@lsi.upc.es

Resumen

Uno de los componentes esenciales de un sistema de extracción de información es el diccionario de patrones necesarios para identificar la información relevante de un documento. Construir un diccionario manualmente además de ser costoso, incide negativamente en la portabilidad del sistema a nuevos dominios. La automatización del proceso de obtención de diccionarios para sistemas de extracción resuelve en parte este problema, aunque sigue precisando la intervención de un experto. En este artículo se propone una metodología para el aprendizaje automático de patrones de extracción partiendo de corpus textuales sin anotaciones, representativos del dominio de trabajo. La metodología incluye diversas etapas, de las cuales destaca la generalización de patrones específicos para obtener patrones de mayor cobertura manteniendo la relevancia de la información extraída. La generalización conlleva además la compactación del diccionario y por tanto reduce el volumen de información a validar por parte del experto.

Palabras clave: Extracción y recuperación de información, aprendizaje automático de patrones de extracción

1 Introducción

El objetivo de un sistema de Extracción de Información (*Information Extraction*) consiste en identificar y extraer información específica de un documento. El tipo de información extraída responde a un conjunto de eventos, entidades y relaciones pre-especificados. A diferencia de un sistema de recuperación de información (*Information Retrieval*), que dada una lista de palabras clave retorna un conjunto de documentos que las contienen, un sistema de EI retorna únicamente la información requerida en un formato prefijado.

Una de las etapas inevitables en la construcción de un sistema de EI es la de la generación de patrones de extracción de información. En los últimos años se han propuesto diversas aproximaciones para resolver esta tarea de forma automática tomando como punto de partida datos que han recibido algún tipo de preprocesamiento, como son los corpus textuales de entrenamiento anotados con etiquetas específicas del dominio, patrones sintácticos y semánticos elaborados manualmente, etc. Esta opción conlleva inconvenientes como son el coste en tiempo destinado a la realización del trabajo manual, que además debe ser llevada a cabo por un experto, y la pérdida de este esfuerzo cuando se intenta trasladar el sistema a nuevos dominios.

Este artículo propone una nueva metodología para la construcción automática de diccionarios de patrones de extracción de información a partir de corpus textuales planos, es decir, sin anotación alguna. La nueva propuesta reduce el esfuerzo del experto humano limitando su intervención a las tareas de validación y tipificación de los patrones de extracción de información, y permite la reutilización de los patrones obtenidos para diversas tareas de extracción.

La siguiente sección del artículo da una breve descripción de la tarea de extracción de información. A continuación, se presenta la problemática de la construcción de diccionarios para sistemas de EI, citándose algunos de los sistemas existentes. La cuarta sección presenta la metodología propuesta en este artículo y se discuten las ventajas que incorpora. Para finalizar se intenta dar una visión futura de la investigación a realizar en la construcción de sistemas de EI "totalmente" automatizados.

2 Extracción de información

La extracción de información es una tarea de procesamiento de lenguaje natural cuyo propósito es extraer determinados tipos de información de un documento. Los sistemas de EI son específicos de un dominio ya que extraen eventos o hechos particulares de un dominio concreto y omiten aquellos que no lo son. Por ejemplo, un sistema de EI en el dominio de noticias sobre fusiones de empresas debería extraer los nombres de las empresas que se fusionan, el capital de la inversión realizada, la actividad o producción de las empresas, su situación geográfica, etc.

En los últimos años se han llevado a cabo diversos congresos centrados en la extracción de información (por ejemplo, los *Message Understanding Con-*

ferences). El objetivo fundamental de estos congresos ha sido la evaluación de sistemas de EI desarrollados en distintos centros de investigación y en cada uno de ellos se ha propuesto un nuevo dominio; por ejemplo, terrorismo en latinoamérica, fusión de empresas o microelectrónica. La organización del congreso proporciona un corpus de entrenamiento y un conjunto de ejemplos (*answer keys*), elaborados manualmente, cuya finalidad es mostrar el tipo de información que el sistema ha de ser capaz de extraer. Las métricas de evaluación están basadas en los valores de dos factores: 1) *recall* que corresponde a la cantidad de extracciones correctas respecto a la cantidad total de extracciones que deberían ser obtenidas y 2) *precision* que indica la cantidad de extracciones correctas respecto a la cantidad total de extracciones efectuadas por el sistema. De la definición de ambos factores se deduce que un intento para mejorar el valor de uno de ellos, incide en el empeoramiento del otro; por esta razón en la evaluación de sistemas de EI se escogen métricas que combinan ambos factores al mismo tiempo, ponderándolos de acuerdo a criterios fijados por la organización.

La construcción de todo sistema de EI requiere un conocimiento considerable sobre el dominio. Por un lado, se requiere conocimiento sobre los objetos del dominio y las relaciones entre estos objetos; por otro lado, se requiere conocimiento sobre cómo se expresan habitualmente en los textos estos objetos y relaciones. A menudo, los objetos del dominio que son relevantes para una tarea de extracción concreta dejan de ser relevantes en otros dominios y tampoco sus relaciones se expresan del mismo modo. La portabilidad de una sistema de EI pasará pues por adquirir este conocimiento, o gran parte de él, de forma automática.

En los sistemas de EI, la forma habitual utilizada para representar contextos locales (que contienen información sintáctica y semántica) necesarios para extraer información es mediante `patrones de extracción`, que en la literatura reciben otros nombres como `reglas de extracción`, `patrones conceptuales` o `nodos conceptuales`. Un patrón de extracción sintetiza el conjunto de restricciones sintácticas y semánticas que debe satisfacer una sentencia para que de ella pueda extraerse información, así como qué elementos de la sentencia serán extraídos. Se dice que un patrón de extracción ha sido activado cuando es aplicable a un fragmento de texto. En este caso se obtiene la información extraída del fragmento, indicada por el patrón.

El conjunto de patrones de extracción de un sistema de EI constituye su `diccionario`. Trasladar un sistema de EI a un nuevo dominio implica construir un nuevo diccionario de patrones de extracción. Si este trabajo se realiza manualmente, requiere mucho esfuerzo por parte de una persona (experta), o equipo de personas, habituadas a este tipo de tarea y con un buen conocimiento del dominio. Además, cuando se desee trasladar el sistema a un nuevo dominio, el trabajo manual deberá repetirse. Una solución a este problema es automatizar el proceso de obtención del diccionario de patrones.

3 Construcción automática de diccionarios de patrones para sistemas de EI

El proceso de obtención de un diccionario de patrones de extracción es uno de los mayores obstáculos a que debe enfrentarse la construcción de un sistema de EI. En los últimos años se han desarrollado diversos sistemas en un intento de resolver esta tarea de forma automática, como son AutoSlog [6] y CRYSTAL [9]. Estos sistemas generan patrones de extracción partiendo de corpus de entrenamiento anotados[1] donde la información a extraer ha sido etiquetada semánticamente[2]. El proceso de anotación de un corpus resulta claramente más sencillo que tener que construir todo un diccionario de patrones, pero aún así es costoso y presenta ciertas dificultades, como decidir qué anotar y cómo hacer la anotación; tampoco evita el tener que recurrir a un experto en el dominio para que lleve a cabo la anotación del corpus.

Otras aproximaciones en la construcción automática de diccionarios de patrones de extracción han conseguido evitar la anotación previa del corpus textual proponiendo soluciones diversas. Por ejemplo, el sistema AutoSlog-TS [8] no requiere corpus anotados pero sí preclasificados, es decir, que los textos que recibe como entrada han sido clasificados como relevantes o irrelevantes de acuerdo al dominio objetivo de la extracción. Con ello, el esfuerzo requerido para obtener un corpus de entrenamiento se reduce considerablemente. Otra propuesta es la que presenta el sistema LIEP [2] que permite que un usuario pueda identificar, de forma interactiva, las entitades de interés y combinaciones de éstas que puedan representar eventos a extraer.

En la primera fase, los patrones de extracción obtenidos presentan una relación muy estrecha con la estructura del corpus de entrenamiento, es decir, reflejan inevitablemente el estilo de redacción usado en él. Si los textos futuros con los que deba trabajar el sistema de EI tienen las mismas características que el corpus de entrenamiento, posiblemente los patrones de extracción serán aún lo suficientemente válidos; en otro caso, para obtener un sistema de EI robusto serán necesarios corpus de entrenamiento de gran tamaño que permitan descubrir todos los posibles patrones necesarios.

Una alternativa mejor es hacer que los patrones iniciales puedan ser generalizados de manera que sean capaces de cubrir ejemplos similares, man-

[1]Otra opción es utilizar un conjunto de *answer keys* que indiquen qué sentencias son relevantes.

[2]Habitualmente, una etiqueta semántica representa el papel que juega el fragmento etiquetado en el contexto en qué se encuentra. Son, por lo tanto, marcas específicas del dominio.

teniendo a su vez cierta especifidad (o sea, con las restricciones necesarias para no cubrir ejemplos que extraerían información irrelevante). Por ejemplo, el sistema CRYSTAL, citado con anterioridad, construye diccionarios de patrones de extracción aplicando un algoritmo similar al de aprendizaje inductivo de conceptos descrito por Michalski [3]. También el sistema LIEP lleva a acabo una generalización de patrones y su punto de vista se acerca más al aprendizaje basado en explicaciones (*EBL*) descrito por Mitchell [5], pero con una teoría del dominio incompleta.

La metodología propuesta en este artículo parte de un corpus textual plano que contiene ejemplos significativos (positivos) del tipo de información que se desea extraer. El corpus inicial es usado como corpus de entrenamiento, del cual se obtienen lo que denominamos "patrones específicos". Estos patrones representan, literalmente, sentencias o partes de sentencias que se hallan en el corpus. Es decir, sólo pueden identificar la información que ellos mismos representan. Para que estos patrones específicos puedan extraer el mismo tipo de información en nuevos textos es necesario someterlos a un proceso de generalización. La generalización de los patrones específicos compacta el diccionario de patrones y facilita el proceso de validación.

La compactación del diccionario mediante la generalización es insuficiente. Muchos de los patrones específicos no podrán ser generalizados y algunos de los patrones generalizados extraerán información irrelevante. De este hecho se deduce que se necesita un mecanismo que determine qué patrones de extracción representan realmente expresiones específicas del dominio: la aplicación de un "filtrado" que deje pasar los patrones con un índice de activación significativo y que la información que extraigan sea relevante, y elimine el resto. La siguiente sección describe la metodología propuesta en este artículo de forma más detallada.

4 Metodología propuesta

Las distintas aproximaciones presentadas en la sección anterior tienen en común la necesidad de trabajar con un experto humano de forma intensiva. La metodología que se propone en este artículo tiene como objetivo principal reducir la intervención y el esfuerzo del experto humano en la tarea de obtención de los patrones de extracción de información. Para conseguir este objetivo la metodología utiliza un algoritmo de aprendizaje o generalización de patrones, y retrasa al máximo la intervención del experto para reducir el volumen de información que debe tratar. El hecho de que el experto intervenga después del proceso de generalización, le permite trabajar directamente con patrones en lugar de trabajar con el corpus. Concretamente, conviene remarcar que partiendo de un corpus sin anotaciones se ahorra mucho trabajo a el experto.

La metodología, expuesta gráficamente en la figura 1, se describe en los siguientes apartados, en los que se detallan brevemente los requisitos, objetivos y funcionamiento de cada una de las fases.

4.1 Obtención de patrones específicos

Esta es la primera fase de la metodología. El punto de partida es un corpus textual sin anotaciones típico del dominio sobre el cual se desea extraer información. El objetivo consiste en obtener un conjunto de patrones específicos que servirán como ejemplos al siguiente proceso de generalización. La obtención de los patrones específicos se inicia con un análisis sintáctico de cada una de las sentencias que componen el corpus de entrenamiento. El análisis sintáctico proporciona los constituyentes sintácticos básicos (sujeto, verbo, objeto directo, etc) que aparecen en una sentencia y sitúa los objetos del texto en estos constituyentes. A partir de este análisis es necesario convertir las sentencias analizadas en patrones específicos traduciéndolas al formalismo escogido para representar los patrones de extracción. Denominamos específico a cada uno de estos patrones porque son patrones de extracción que solamente cubren la sentencia que representan. A continuación se muestra un ejemplo sencillo de patrones específicos obtenidos a partir de dos sentencias.

```
Sentence:  At least one person has been killed in an avalanche
in the Italian Alps.

        Pattern <kill-passive1>
        subject constraints = one person
        verb constraints = kill passive
        pp constraints = Italian Alps
        pp constraints = avalanche

Sentence:  Five people were killed this weekend in helicopter
crash in the Rocky Mountains of Southeastern B-C.

        Pattern <kill-passive2>
        subject constraints = Five people
        verb constraints = kill passive
        pp constraints = helicopter crash
        pp constraints = Rocky Mountains of Southeastern B-C
        pp constraints = this weekend
```

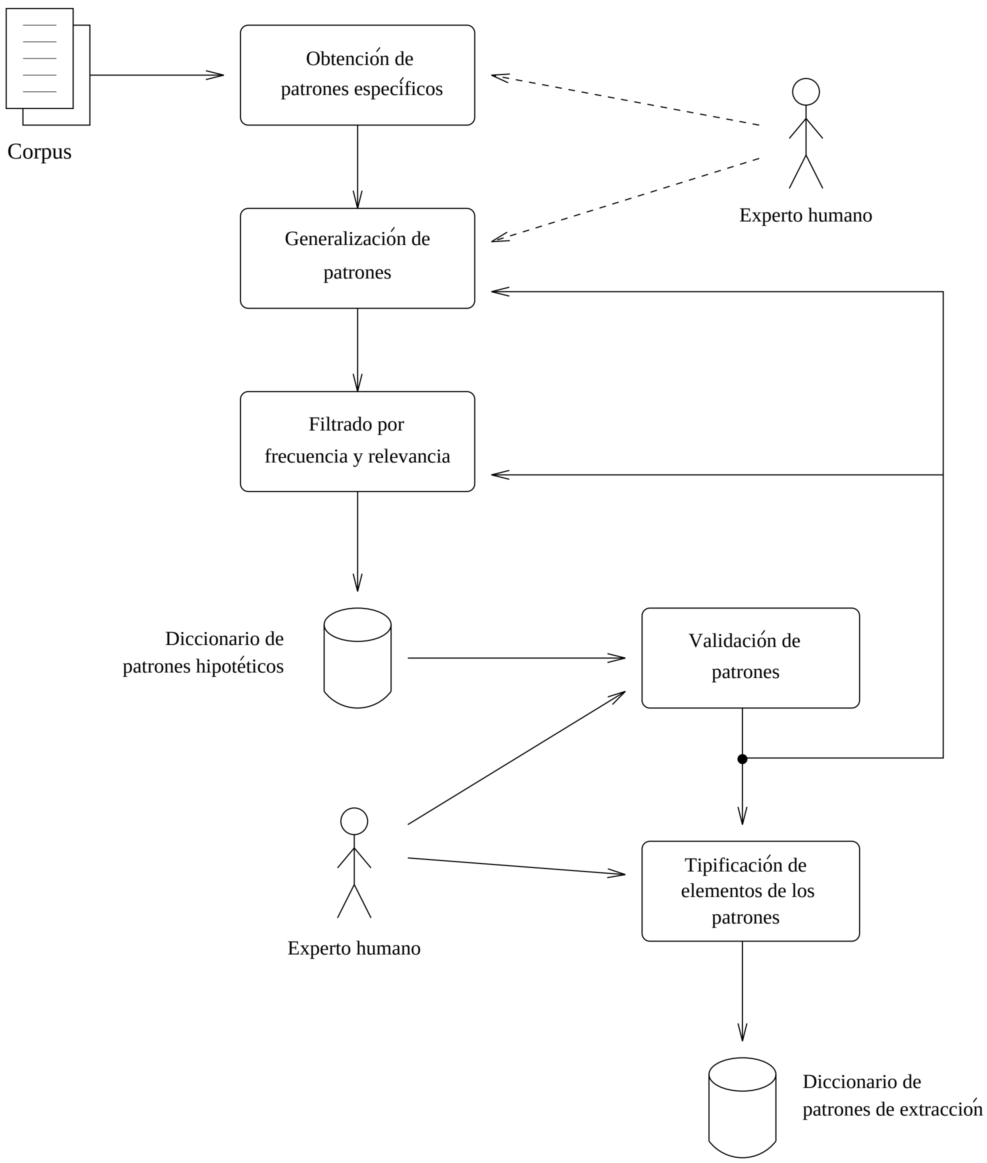

Figura 1: *Organigrama de la metodología.*

Opcionalmente, el experto puede proporcionar información (pistas) sobre el tipo de sentencias que son relevantes para obtener patrones específicos. Por ejemplo, podría dar un léxico de palabras clave que deberían aparecer en las sentencias como condición para ser consideradas relevantes. El léxico dado por el experto no tendría por qué se exhaustivo sino que podría ser ampliado automáticamente con la ayuda de un tesaurus (e.g. WordNet [4]) que proporcionase sinónimos de las palabras clave. Por ejemplo, en el dominio de actos terroristas, el experto indicaría que la palabra "kill" es relevante sin tener que indicar que también los son otras palabras clave sinónimas que se obtendrían usando WordNet:

kill : eliminate, annihilate, extinguish, eradicate, wipe out, carry off, cancel out, drown, massacre, slaughter, mow down, butcher, slaughter, chine, poison, stone, lapidate, poison, brain, put away, put to sleep, liquidate, waste, knock off, do in, dispatch, exterminate, kill en masse, kill off, smother, asphyxiate, suffocate, strangle, throttle, garrotte, garotte, decapitate, behead, guillotine, impale, stake, dismember, cut to pieces, tear to pieces, quarter, draw and quarter, hang, string up, murder, slay, hit, dispatch, bump off, polish off, remove, burke, execute, murder execution-style, assassinate, execute, put to death, crucify, kill by crucifixion, electrocute, fry, burn, burn at the stake, hang, lynch, shoot, pick off, shoot one by one.

A pesar de que en la lista de sinónimos obtenida a partir de WordNet puedan aparecer palabras no relevantes para el dominio de actos terroristas, no es ningún inconveniente ya que no habrán sentencias que las contengan o si las hay, su escasa frecuencia de aparición hará que acaben eliminándose en los procesos de filtrado.

4.2 Generalización de patrones

Un patrón de extracción específico reconoce una sentencia concreta del texto, es decir, que utiliza las mismas palabras que encuentra en el texto para representar lo que serán las restricciones de aplicabilidad del patrón. A priori resulta difícil saber qué características del patrón son fundamentales para su propósito de extracción y cuales no lo son. La idea esencial del proceso de generalización consiste en obtener patrones de extracción capaces de extraer la misma información que podía extraerse con los patrones específicos, sin que se extraiga información irrelevante y, al mismo tiempo, se reduzca considerablemente el volumen inicial de patrones.

El proceso automático de generalización de patrones se basa pues, en el conjunto de patrones específicos obtenidos en la primera etapa. El algoritmo trata de encontrar generalizaciones que cubran diversos patrones específicos, es decir, que a partir de un patrón generalizado se pueda extraer la misma información que la obtenida a partir de los patrones específicos que decimos que cubre.

La generalización se lleva a cabo a partir de elementos de los patrones que llamaremos "generalizables", como pueden ser:

- **Contenido semántico de los constituyentes sintácticos:** Consiste en hacer la generalización a partir de los rasgos semánticos asociados a los elementos que forman parte de los constituyentes sintácticos.

  Para llevarla a cabo es necesario disponer, además del léxico, de un mecanismo que nos proporcione el rasgo semántico asociado a cada elemento así como una jerarquía de rasgos semánticos. También para esta tarea resulta de gran utilidad WordNet.

  La generalización de un constituyente sintáctico, que en un patrón tiene como rasgo semántico A y en otro B, sería [A $\vee$ B .. T]. El significado de esta expresión es que el patrón generalizado resultante tiene como posibles rasgos semánticos A, B y los hiperónimos de A y de B que se encuentren por debajo de T, siendo T el primer nodo de la jerarquía semántica que subsume a ambos rasgos.

- **Presencia de constituyentes sintácticos:** La idea es eliminar de los patrones, aquellos constituyentes sintácticos que diferencian a un conjunto de patrones muy similares. De este modo, el patrón resultante cubrirá todos los patrones similares omitiendo características que dada su gran diversidad parecen irrelevantes.

A continuación se muestra la generalización obtenida de los dos patrones específicos vistos anteriormente.

```
Generalized Pattern <kill-passive>
subject constraints = [person]
verb constraints = kill passive
pp constraints = [location]
```

En este caso, "person" posee el rasgo semántico "person" y "Five people" posee el rasgo semántico "people", ambos obtenidos de WordNet. Mediante una generalización del contenido semántico del constituyente sujeto, obtenemos valores semánticos en el rango [person $\vee$ people .. person], ya que "person" es el primer nodo en la jerarquía semántica que subsume a ambos. La simplificación del rango nos lleva a [person]. Del resto de constituyentes sintácticos sólo el que posee como rasgo semántico "location" ("Italian Alps" y "Rocky Mountains of Southeastern B-C" tienen este mismo rasgo semántico) es compartido por los dos patrones específicos. Y por tanto, tras la eliminación de constituyentes sintácticos no comunes es el único que se mantiene.

El algoritmo de generalización utilizado será incremental, para permitir el tratamiento de nuevos documentos del mismo dominio sin tener que rehacer el trabajo hecho. También permitirá el tratamiento de ejemplos negativos, pues como veremos más adelante, el experto frente a un patrón generalizado que extrae información irrelevante, puede marcar los ejemplos del corpus incorrectamente cubiertos por el patrón generalizado como ejemplos negativos.

En el proceso de generalización, junto a la estructura del patrón general que se va construyendo, se mantiene el conjunto de las sentencias del corpus que cubre. Así pues, para cada patrón, además de disponer de su descripción tenemos el conjunto de ejemplos del corpus que cubre.

Opcionalmente, el experto puede proporcionar información para guiar el aprendizaje indicando qué elementos deberían formar parte de un patrón. Por ejemplo, su conocimiento del dominio le permite señalar que en el dominio de trabajo determinadas formas verbales son altamente indicativas del tipo de información que se desea extraer. En la literatura del aprendizaje automático ésto puede verse como un sesgo (*bias*) para acelerar el proceso de aprendizaje.

4.3 Filtrado de patrones

La cantidad de patrones de extracción obtenidos puede seguir siendo muy elevada incluso después del proceso de generalización. Algunos de ellos resultarán poco útiles para el proceso de extracción de información relevante del dominio y habrá otros de espurios. Para tratar de resolver este problema, los patrones obtenidos después del proceso de generalización son sometidos a un proceso de filtrado, cuyo objetivo es precisamente eliminar aquellos que sean espurios o irrelevantes.

Los posibles procesos de filtrado a aplicar serán:

- **Filtrado por frecuencia:** Este proceso intenta eliminar patrones generales que tienen una aplicabilidad muy reducida. Para eliminarlos, se fija un umbral y en el caso de que el número de aplicaciones posibles de un patrón en el corpus no supere dicho umbral, será considerado como un patrón espurio y será eliminado. La idea es pues que, mediante este proceso, consigamos quedarnos únicamente con patrones mínimamente útiles.

- **Filtrado por relevancia:** Este proceso solamente puede llevarse a cabo cuando se dispone de ejemplos de extracción irrelevante o poco satisfactoria. Como se explicará a continuación, el experto encargado de supervisar los patrones obtenidos puede indicar qué ejemplos cubiertos por un patrón general no deberían estarlo. Por lo tanto, propone que el patrón ha sido incorrectamente generalizado o bien que la información que extrae es irrelevante.

El filtrado por relevancia consiste en determinar qué patrones de extracción son realmente relevantes y en eliminar aquellos que no superen un umbral prefijado de relevancia. El cálculo de la relevancia de un patrón de extracción corresponde al número de ejemplos relevantes que cubre respecto al número total de ejemplos que cubre.

4.4 Validación de patrones

En este punto, el volumen de patrones iniciales se ha reducido considerablemente y por lo tanto el coste del proceso de revisión también. Obtenido ya un conjunto de patrones hipotéticos, éste debe ser validado por un experto. Para ello, el experto examina cada uno de los patrones obtenidos y da su opinión. Si decide que el patrón no es adecuado, por ser demasiado general, puede indicar qué ejemplos de los que van asociados al patrón no deberían ser cubiertos. Estos ejemplos de recuperación indebida sirven de ejemplos negativos en una nueva aplicación del algoritmo de generalización.

En la validación, si el experto considera que la cantidad de patrones obtenidos es excesiva, puede modificar los parámetros y métodos de filtrado así como la forma de generalizar proponiendo un nuevo sesgo más adecuado. El procedimiento se repite hasta que el experto considere que los patrones son satisfactorios, momento en que se pasa a la siguiente fase.

En el caso de la generalización del patrón visto anteriormente, si el experto considera que el patrón es demasiado general, ya que por ejemplo no permite extraer el momento del accidente, puede marcar los patrones específicos que cubre el patrón indebidamente (<kill-passive2>) como ejemplos negativos para ese patrón y repetir el proceso de generalización.

4.5 Tipificación de patrones

Hasta esta fase, una vez validado el diccionario de patrones hipotéticos, el experto no ha determinado aún el tipo concreto de información que se desea extraer de cada patrón.

La tipificación consiste en dar nombres, que en realidad son las funciones que desempeñan, a los distintos elementos del patrón indicativos del tipo de información que extraerán. Por ejemplo, volviendo al dominio de actos terroristas, en un patrón que represente el conjunto de sentencias “[person]

was assassinated", tipificará [person] como [VICTIM]. [VICTIM] es la función que desempeña [person] en el patrón y representa el tipo de información que desea extraerse en este dominio.

Todas las aproximaciones presentadas en este artículo llevan a cabo la tipificación de forma manual. Ya sea partiendo de corpus con anotaciones semánticas o *answer keys*, proporcionando un sistema interactivo para que el experto defina los eventos o bien tipificando los patrones obtenidos a posteriori con ayuda del experto, el problema es el mismo: hay que recurrir a un experto para que realice esta tarea. En el siguiente apartado, proponemos una alternativa a la tipificación manual de patrones a estudiar en trabajos futuros.

Una ventaja de la tipificación tardía de los patrones de extracción, añadida a la reducción del volumen de información a tratar, es que favorece su reutilización para otras tareas relacionadas con la extracción. Por ejemplo, la clasificación de textos [7], la obtención de resúmenes, la construcción de léxicos específicos que incorporan información contextual o la construcción de herramientas destinadas a la desambiguación léxica.

5 Conclusiones y futuros trabajos

Este artículo propone una metodología para la construcción automática de diccionarios de patrones para sistemas de extracción de información. Con ella se pretende: 1) evitar el esfuerzo requerido en la preparación de un corpus textual de entrenamiento y 2) reducir la intervención del experto humano en el proceso de obtención de patrones generales de extracción.

La metodología comprende cinco etapas básicas, a las cuales pueden ser añadidas diversas opciones para guiar su estrategia. A partir de la primera etapa se obtiene un conjunto de patrones específicos que sirven de ejemplos para la siguiente etapa de aprendizaje o generalización. La generalización no garantiza la relevancia de los patrones ni tampoco su grado de utilidad, por este motivo los patrones son sometidos a un proceso de filtrado. La relevancia de un patrón va ligada al propósito de la extracción y quien determina qué se extraerá, es un experto. Así, es el experto el encargado de validar los patrones, y de someterlos a una nueva generalización si es necesario, y de tipificarlos (indicar el tipo de información que será extraída).

El propósito de los patrones de extracción es el de obtener determinada información de un documento. Como hemos visto, qué extraerá concretamente un patrón no se decide hasta que éste sea tipificado. Pero la tipificación no es más que la identificación de un elemento de un patrón con la función que representa este elemento en su contexto (los elementos que le rodean y que también forman parte del patrón). Determinar esa función requiere mucho conocimiento pero no es imposible obtenerla si se dispone de un sistema que permita expresar el concepto que representa. La representación conceptual de la función unida a un sistema que permita clasificar instancias (en este caso patrones) bajo el concepto adecuado, podrían ser suficientes para conseguir automatizar el proceso de tipificación. Un sistema que responde perfectamente a estas características es YAYA [1] y con él intentaremos implementar un proceso automático de tipificación.

Referencias

[1] Jordi Àlvarez. Yet another yet another (YAYA). Technical Report LSI-96-15-T, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, 1996.

[2] Scott B. Huffman. Learning information extraction patterns from examples. In *IJCAI-95 Workshop on New Approaches to Learning for NLP*, 1995.

[3] R.S Michalski. A theory and methodology of inductive learning. *Artificial Intelligence*, 20:111–161, 1983.

[4] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to wordnet: An on-line lexical database. Technical report, 1993.

[5] T.M. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.

[6] Ellen Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 1993.

[7] Ellen Riloff. Using learned extraction patterns for text classification. In G. Scheler S. Wermter, E. Riloff, editor, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 275–289. Springer-Verlag, 1996.

[8] Ellen Riloff and Jay Shoen. Automatically acquiring conceptual patterns without and annotated corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 148–161, 1995.

[9] Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy Lehnert. Crystal: Inducing a conceptual dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995.