Un sistema de extracción de información sobre desastres naturales

Alberto Téllez Valero[1], Manuel Montes-y-Gómez[1,2], Luis Villaseñor Pineda[1]

[1]Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica, México
[2]Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, España
{albertotellezv, mmontesg, villasen}@inaoep.mx

Resumen. Hoy día el acceso a vastas cantidades de información dificulta su exploración y análisis de forma manual. Una manera de confrontar el problema es con *extracción de información*, donde la tarea es filtrar y estructurar de manera automática textos en lenguaje natural. En este trabajo se describe el sistema *Topo*, un sistema de extracción de información que toma como dominio de estudio noticias que reportan desastres naturales en español. Se presenta la arquitectura del sistema y los resultados alcanzados actualmente tanto en el filtrado de textos relevantes al dominio, así como en la identificación y extracción de entidades de información. Además, se muestran las conclusiones y el trabajo en proceso.

1 Introducción

El crecimiento explosivo de documentos en lenguaje natural disponibles en computadoras conectadas a la red alrededor del mundo dificulta su exploración y análisis. Por tal motivo, se hace necesario poder filtrar y estructurar información relevante al dominio de interés para contestar a muchas preguntas acerca del mismo. Una forma de lograr esto último de manera automática es con *extracción de información* (EI). La extracción de información es la tarea de localizar piezas específicas de datos desde documentos en lenguaje natural [1], la información extraída es entonces almacenada en una base de datos que puede ser examinada usando lenguajes de consultas estándar para facilitar su análisis.

En este artículo presentamos un sistema que realiza la tarea de extracción de información para el dominio de noticias que tratan sobre desastres naturales. El objetivo principal de nuestro sistema, al que hemos denominado *Topo*, es mostrar el uso de técnicas de clasificación automática de textos en la tarea de extracción de información. Además, se pretende mostrar que para la extracción de ciertas entidades de información es suficiente con emplear un análisis a nivel de palabras, sin tener que lidiar con el hasta hoy incompleto análisis lingüístico, especialmente para el idioma español. Esto último implica el poder extraer información sin tratar de comprender el contenido del texto, que es una postura contraria a lo que generalmente se piensa.

A. Gelbukh, G. Sidorov, W. Olán, J. Vera (Eds.) Recientes avances en la ciencia de la computación en México, pp.89-98.

El resto del documento se encuentra organizado de la siguiente manera. En la sección 2 describimos brevemente el trabajo relacionado. En la sección 3 se presenta el dominio de trabajo y la plantilla de extracción a llenar. La sección 4 exhibe la arquitectura del sistema. Los resultados preliminares se muestran en la sección 5. Finalmente, en la sección 6 se presentan las conclusiones y se describe el trabajo en proceso.

2 Trabajo Relacionado

Un auge importante en el desarrollo de sistemas de extracción de información se dio gracias a la intervención de la Agencia de Defensa de los Estados Unidos (DARPA, por sus siglas), quien fomentó las Conferencias de Entendimiento de Mensajes (MUC, por sus siglas en inglés), las cuales proporcionaron más de una década de experiencia en la definición, diseño y evaluación de este tipo de sistemas [2]. Entre otras cosas, los resultados del MUC demostraron que la extracción de información es una tarea difícil hasta para las personas, donde se reportó que los humanos podemos alcanzar un grado de exactitud entre el 60 y 80% en esta labor [1].

Generalmente, la mayoría de sistemas de extracción de información se basan en arquitecturas como la propuesta por Grishman [3]. En esta arquitectura se tiene una estructura modular, donde la salida de un módulo sirve como entrada del módulo siguiente. Una característica importante de la misma es el amplio uso de recursos lingüísticos para las tareas de análisis léxico, análisis sintáctico parcial y resolución de correferencia. Esto se debe principalmente a que se tiene la hipótesis que para lograr extraer información de un texto se debe entender en el mayor grado posible el mismo.

En contraste a la arquitectura de Grishman, Kushmerick *et al* [4] proponen una arquitectura novedosa, donde la hipótesis no es entender el texto, si no más bien encontrar las combinaciones de palabras (expresiones) que se utilizan para reportar la información que nos interesa extraer. La ventaja de esta arquitectura es que no se tiene que lidiar con un profundo análisis lingüístico, en su lugar se utilizan técnicas de clasificación automática de textos para encontrar las expresiones buscadas. Esta arquitectura fue presentada en el sistema llamado "Cambio de Dirección" (CoA, por sus siglas en inglés), el cual tiene como objetivo filtrar correos electrónicos que reportan un cambio de dirección de e-mail por parte del remitente, y posteriormente actualizar la agenda del destinatario con la información incluida en los mismos. Sin embargo, extraer direcciones de e-mail no refleja completamente los alcances de la arquitectura, por tal motivo nosotros pretendemos llevar a una tarea más compleja este trabajo, con el propósito de tratar de determinar sus límites.

Con respecto a extraer información de noticias que reportan desastres naturales, actualmente no tenemos antecedentes de la existencia de sistemas de extracción de información enfocados al dominio, sólo sabemos que existe una asociación que realiza esta tarea de forma manual (consultar referencia [5]).

3 Dominio

La información a ser extraída se define por medio de la plantilla de extracción, la cual se forma por una serie de atributos que la caracteriza. Los atributos pueden ser opcionales u obligatorios ya que la información puede o no estar presente en los documentos. Su construcción se realiza de antemano y dependen del dominio de trabajo, también llamado escenario, y de la información que se desea obtener.

En nuestro caso, el escenario en el que se decidió trabajar es el de noticias en español que reportan desastres naturales. La razón de su elección es que es un dominio rico en información para ser extraída. No obstante, la principal motivación para esta elección fue nuestra convicción respecto a que la disponibilidad de un inventario con dicha información, en combinación con un conjunto de herramientas para su adecuado análisis, permitirá adquirir un mejor conocimiento sobre los fenómenos naturales desastrosos, y con ello aprender a prevenir y minimizar sus efectos.

Debido a que son muchos los tipos de desastres naturales que se presentan, actualmente sólo nos enfocamos en seis de los más frecuentes en México (ver tabla 1). Las definiciones incluidas en la tabla 1 corresponden con las establecidas en la Guía Metodológica de DesInventar publicada en el 2003 [5], esta guía fue elaborada por la Red de Estudios Sociales en Prevención de Desastres en América Latina (LA RED). Entre otras cosas, DesInventar presenta una metodología de registro de información sobre características y efectos de diversos tipos de eventos. Por tal motivo, los datos a extraer en el presente proyecto, también corresponden con los establecidos por LA RED en su Ficha de Información de Desastres. La plantilla de extracción se muestra en la tabla 2.

Tabla 1. Dominio de estudio

Desastre	Definición
Helada	Disminución de la temperatura hasta el punto de congelación con efectos nocivos en la población, cultivos, bienes y servicios
Huracán	Anomalía atmosférica violenta que gira a modo de torbellino caracterizado por fuertes vientos, acompañados por lluvia
Forestal	Incendio. Incluye todos los incendios en campo abierto en áreas rurales, sobre bosques nativos, bosques cultivados y praderas
Inundación	Subida de aguas que supera la sección del cauce de los ríos o que se relaciona con el taponamiento de alcantarillas
Sequía	Temporada anormalmente seca, sin lluvias, o con déficit de lluvias. En general se trata de períodos prolongados
Sismo	Todo movimiento de la corteza terrestre que haya causado algún tipo de daño o efecto adverso sobre comunidades o bienes

Tabla 2. Plantilla de extracción

Relacionados con el desastre	
Fecha	Fecha de ocurrencia del desastre
Lugar	Nombre del lugar o lugares donde ocurrió el fenómeno
Magnitud	Valores de magnitud internacionalmente usados para sismo y huracán, para otros tipos de eventos variables cuantificadas
Relacionados con las personas	
Muertos	Número de personas fallecidas por causas directas
Heridos/ Enfermos	Número de personas que resultan afectadas en su salud o integridad física, sin ser víctimas mortales, por causa directa del desastre
Desaparecidos	Número de personas cuyo paradero a partir del desastre es desconocido
Damnificados	Número de personas que han sufrido grave daño directamente asociados al evento en sus bienes o servicios
Afectados	Número de personas que sufren efectos secundarios asociados a un desastre
Relacionados con las viviendas	
Destruidas	Número de viviendas arrasadas, sepultadas, colapsadas o deterioradas de tal manera que no son habitables
Afectadas	Número de viviendas con daños menores, no estructurales o arquitectónicos, que pueden seguir siendo habitadas
Relacionados con la infraestructura	
Vías	Longitud de redes viales destruidas o inhabilitadas en metros
Hectáreas	Número de áreas de cultivo, pastizales o bosques destruidas o afectadas
Ganado	Número de unidades perdidas (bovinos, porcinos, caprinos, avícolas)
Centros educativos	Número de guarderías, colegios, universidades, centros de capacitación, etc. destruidas o afectadas directa o indirectamente por el desastre
Centros de salud	Número de centros de salud, clínicas, hospitales destruidos o afectados directa o indirectamente por el desastre
Perdida económica	Monto de las pérdidas directas causadas por el desastre

4 Arquitectura

Básicamente, para realizar extracción de información bajo la arquitectura presentada por Kushmerick *et al* [4], se requiere en primer lugar filtrar los textos relevantes al dominio de estudio, posteriormente detectar las entidades con posibilidad de ser extraídas, y finalmente discriminar entre las entidades identificadas las que proporcionan información útil para llenar la base de datos. El resultado final es una base de datos constituida por la colección de plantillas generadas a partir de los textos filtrados.

En la figura 1 se muestra la interfaz del sistema *Topo*, en esta figura se presenta el ejemplo de una noticia que reporta un "sismo", y con la certeza de que es un texto relevante al dominio, el sistema se ocupa de identificar las entidades con posibilidad de ser extraídas (las frases sombreadas en el cuadro de texto *Noticia*). Para posteriormente, discriminar entre estas entidades aquellas que son útiles para llenar la plantilla de extracción. Para efectuar esto último, a cada una de las entidades se les calcula la probabilidad de que formen parte o no en los registros de la plantilla (cuadro de texto *Distribución*), y se toman como resultado las que tienen una mayor probabilidad. De la información extraída en el ejemplo (panel *Plantilla*), podemos concluir que se reporta un sismo de 7.6 grados que tuvo lugar en Colima, México, dejando un saldo de 28 muertos, 300 heridos o enfermos y 10 mil viviendas afectadas. Lo anterior refleja la utilidad de este tipo de sistemas.

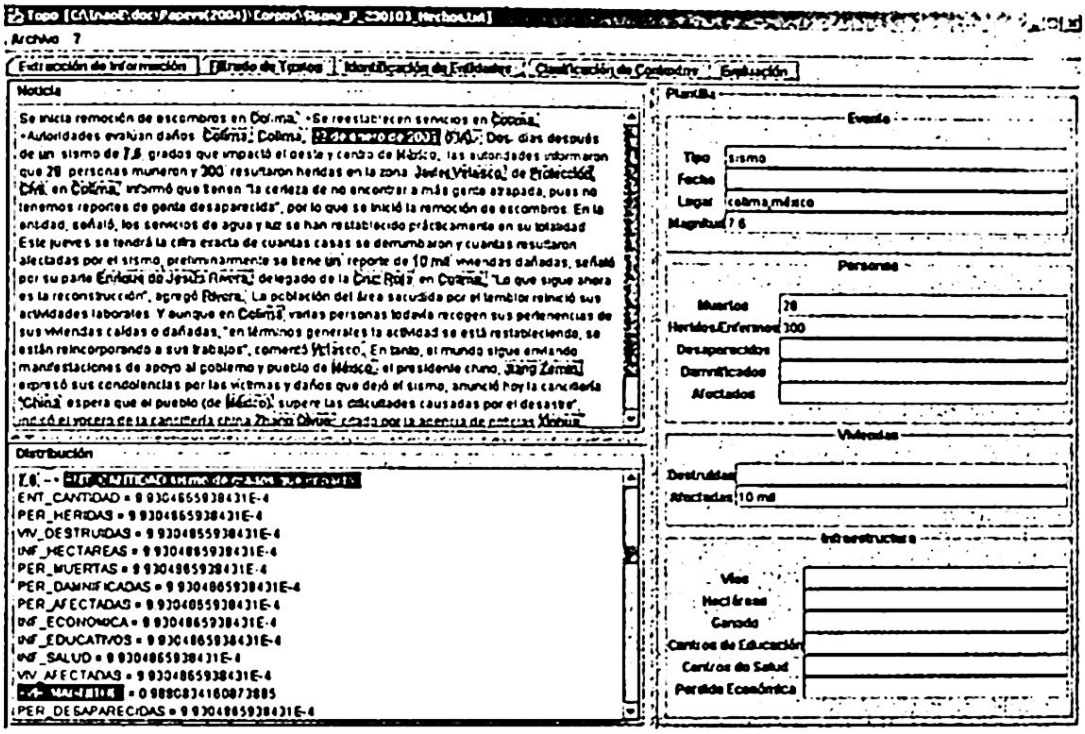

Fig. 1. Interfaz del sistema

El diseño arquitectónico del sistema *Topo* se muestra en la figura 2, este diseño consiste de un modelo estructural compuesto por tres componentes de programa. Donde, la salida del primer componente es un condicionante para la ejecución de los dos siguientes, y la salida del segundo componente sirve como entrada del último. Además, cabe resaltar que el trabajo de extracción recae principalmente en un análisis a nivel de palabras y el uso de clasificadores de textos. Estas características proporcionan flexibilidad al sistema para ser adaptado a nuevos dominios. En las subsecciones siguientes se explica más a detalle cada uno de los componentes de la arquitectura.

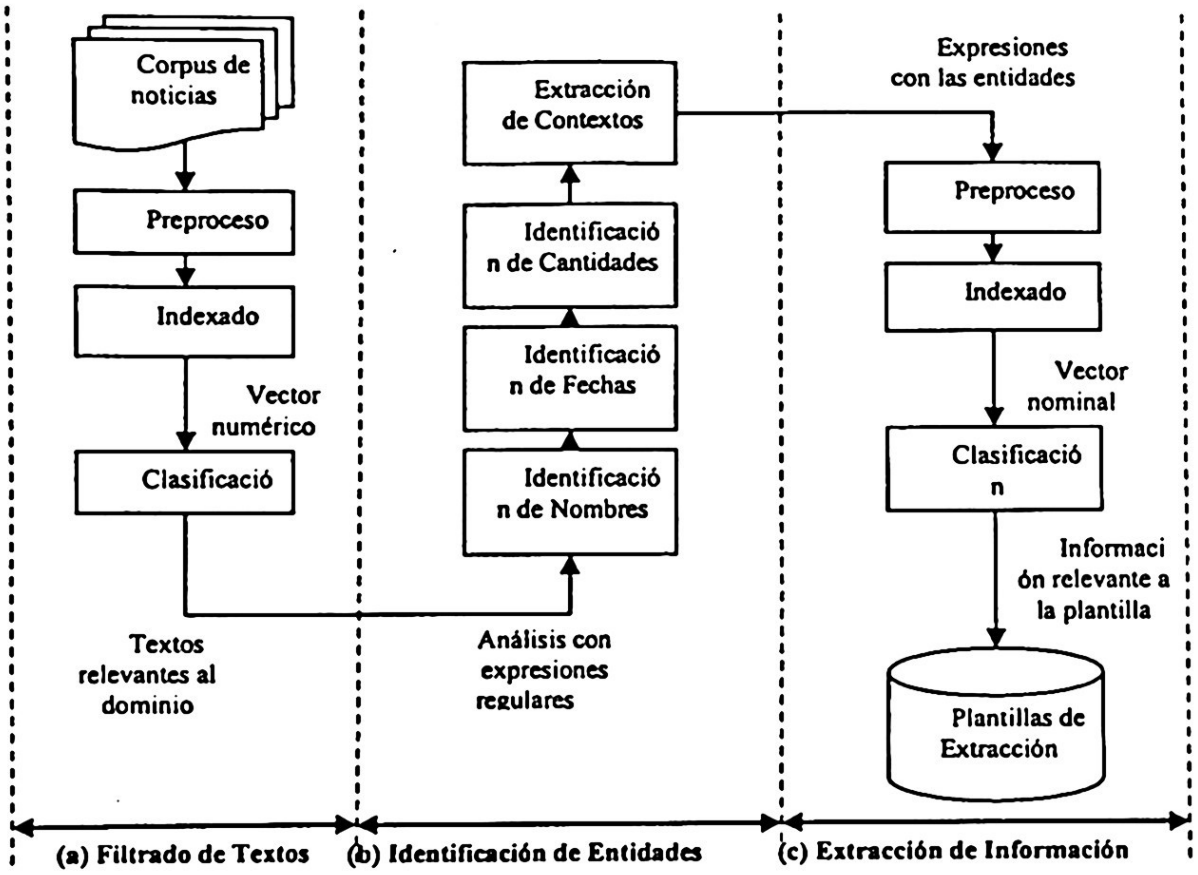

Fig. 2. Arquitectura del Sistema

4.1 Filtrado de texto

El objetivo de este componente es filtrar aquellos documentos que son relevantes al dominio de trabajo. Una forma efectiva de realizar esta tarea es utilizando técnicas de *clasificación automática de textos* (CT). Donde la clasificación de textos se define como la tarea de ubicar correctamente y de manera automática textos en lenguaje natural que contienen información no estructurada en un conjunto de categorías predefinidas [6]. Para la construcción del clasificador se utilizaron técnicas de aprendizaje automático (para más detalle ver referencia [7]). Generalmente, el proceso de clasificar un texto consiste en:

1. Realizar la extracción de características para transformar el texto de su formato inicial a una representación adecuada para la tarea de clasificación.
2. Aplicar el método de clasificación.

En nuestro caso, el modelo utilizado para la representación fue el vectorial junto con un indexado booleano [6]. En otras palabras, el documento está representado por un vector de 0's y 1's que indican la presencia o ausencia de ciertas palabras en el mismo. Para mejorar el indexado se aplicó un preproceso al texto donde se eliminaron símbolos de puntuación (comas, puntos, etc.). Con respecto al método de clasificación, el que se está utilizando actualmente es el simple de Bayes (ver sección

4.3 para más detalle), el cual fue entrenado con un corpus de 471 noticias, de las cuales el 48% es relevante y el 52% restante es irrelevante[1].

Una ventaja de utilizar un clasificador de textos es que el componente puede ser fácilmente adaptado a un nuevo dominio. Para esto, sólo es necesario contar con un conjunto de documentos previamente clasificados del dominio de interés, esto para entrenar nuevamente el método de clasificación.

4.2 Identificación de entidades

Este componente se encarga de detectar aquellas partes del texto con posibilidad de ser incluidas en la plantilla de extracción. En este caso, las entidades probables son: *Nombres* que puedan reflejar el lugar del evento; *Fechas* que puedan reportar la ocurrencia del mismo, y *Cantidades* relacionadas con la magnitud del evento o con efectos sobre personas, viviendas o infraestructura.

Para identificar las entidades se utiliza un análisis con expresiones regulares a partir de los componentes léxicos de la gramática siguiente:

Ent_nombre	→	**nombre | nombre con_nombre** ent_nombre
ent_fecha	→	**mes | mes con_fecha número | número con_fecha** ent_fecha
ent_cantidad	→	**número(. número)? | número(. número)?** ent_cantidad

Donde los terminales generan conjuntos de cadenas dados por las siguientes definiciones regulares:

nombre	→	[A-Z][A-Za-z]*
con_nombre	→	(de | la | ... | se)*
mes	→	enero | ... | diciembre
con_fecha	→	de | - | ... | ϵ
número	→	[0-9]+

Además, a las definiciones regulares *nombre* y *número* se les agrego, respectivamente, un diccionario de expresiones que representan excepciones a la definición. Por ejemplo, las palabras al principio de una oración que inician con letra mayúscula y que no son nombres propios. (*El* sismo ...), y los números reportados con letras, o una combinación de letras y números (... dejó *mil 500* muertos).

Este tipo de análisis nos ha resultado en una baja precisión para identificar nombres, pero en una alta cobertura para todos los casos. La completa cobertura es importante para no dejar fuera del proceso de extracción entidades con alguna probabilidad de formar parte en la plantilla. Finalmente, después de identificar las

[1] Por textos relevantes entendemos todos aquellos que contienen información a ser extraída, mientras que los irrelevantes son los que contienen palabras o frases usadas comúnmente en la descripción de un fenómeno natural, pero que en estos casos se usan en contextos muy diferentes. Por ejemplo, la frase "ojo del huracán" en el contexto de "el presidente está en el ojo del huracán".

entidades, este componente se encarga de extraer los contextos de cada entidad, donde el contexto de la entidad es la expresión donde se encuentra la misma.

También, cabe mencionar que el tipo de entidades identificadas actualmente pueden ser útiles para detectar información en otros dominios, y no exclusivamente del escenario actual, por lo que este componente es completamente adaptable a tareas de extracción similares.

4.3 Extracción de información

Para filtrar entre las entidades identificadas aquellas que son útiles a la plantilla, tomamos a la tarea de extracción de información como una tarea de clasificación de textos, con la variante de que lo que estamos clasificando es el contexto de las entidades identificadas en lugar de los textos completos. Esta clasificación es la que nos permite conocer la probabilidad de que la entidad forme parte en alguno de los registros de la plantilla, o bien que no sea tomada en cuenta.

Para hacer la clasificación de un contexto, la parte de extracción de características es similar a la explicada en la tarea de filtrar un texto (sección 4.1), con la única diferencia que en el indexado el vector resultante tiene como entradas atributos nominales (palabras del contexto) y no atributos numéricos (ponderado Booleano). Finalmente, para discriminar los contextos se usan tres clasificadores de textos, los cuales fueron entrenados con una colección de 2,364 expresiones formadas de seis palabras, las cuales fueron obtenidas de 90 documentos que tratan sobre incendio forestal y sismo. En las expresiones el 26% representan información útil para la plantilla, y el 74% restante son entidades identificadas pero que no deben ser extraídas.

El uso de tres clasificadores se debe a que se especializó cada uno de ellos en el contexto de nombres, fechas y cantidades respectivamente. Los algoritmos utilizados hasta el momento son el simple de Bayes para las fechas y vecinos más cercanos para los otros dos casos.

El método simple de Bayes es del tipo probabilístico y es construido utilizando el conjunto de entrenamiento para estimar la probabilidad de cada clase dadas las características de los textos. Para evaluar dicha probabilidad se utiliza una simplificación del teorema de Bayes:

$$P(c_j \mid d) = P(c_j) \prod_{i=1}^{M} P(d_i \mid c_j) \tag{1}$$

Donde M es el número de términos en el modelo vectorial del texto d_i y $1 \leq j \leq k$, donde k es el número de clases posibles. Las probabilidades $P(c_j)$ y $P(d_i \mid c_j)$ se calculan de la siguiente manera:

$$P(c_j) = \frac{N_j}{N} \qquad P(d_i \mid c_j) = 1 + N_{ij} \Big/ M + \sum_{k=1}^{M} N_{kj} \tag{2}$$

Aquí N es el número de documentos en el conjunto de entrenamiento, N_j es el número de documentos en el conjunto de entrenamiento con clase c_j, y N_{ij} es el número de veces que la palabra i ocurre dentro de los textos con clase c_j en el conjunto de entrenamiento.

Con respecto al método de vecinos más cercanos, este es un método de aprendizaje basado en instancias que consiste en almacenar los datos de entrenamiento para que dado un nuevo texto, se busque en los datos almacenados un caso similar y se clasifique en base a la clase de ese ejemplo similar. Se usa una función de distancia para determinar cual miembro del conjunto de entrenamiento es el más cercano al nuevo caso. La función de distancia más usada es la distancia Euclidiana [6].

Un aspecto importante a resaltar es que el uso de clasificadores de textos permite adaptar el componente a nuevos dominios. Para entrenar nuevamente los clasificadores es necesario contar con una colección de contextos de las entidades nombradas para el nuevo escenario, y además indicar cuales de las expresiones representan información útil a la plantilla y que registro les corresponde.

5 Resultados preliminares

Para el proceso de evaluación utilizamos las medidas de precisión y cobertura, donde el objetivo es valorar la respuesta del sistema basándose en las decisiones del experto. En la evaluación del filtrado de textos y la extracción de información se utilizó el método de validación cruzada con 10 pliegues (10FCV, por sus siglas en inglés) [8]. Los resultados alcanzados hasta el momento se muestran en la tabla 3. Cabe mencionar que los clasificadores incluidos en el componente de extracción de información actualmente han sido entrenados con contextos de tamaño seis (tres palabras a la izquierda y tres palabras a la derecha para cada entidad), y aún no estamos seguros que este tamaño sea el mejor para todos los casos, por tal motivo los resultados se consideran preliminares.

Tabla 3. Resultados Preliminares

Componente	Cobertura	Precisión
Filtrado de textos	96%	96%
Identificación de entidades	99%	88%
Extracción de información	71%	72%

6 Conclusiones

En base a los resultados preliminares podemos concluir que es posible hacer extracción de información de entidades que están de forma explícita en los textos utilizando únicamente un análisis a nivel léxico junto con métodos de clasificación de textos. Esta conclusión es importante porque se muestra que para tareas como la presentada en este trabajo, no es necesario tener un amplio entendimiento del lenguaje para lograr extraer información, que es lo que generalmente se intenta. Además, esto fue probado para un dominio más complejo que el presentado por Kushmerick et al [4].

Hasta el momento, el principal problema que hemos afrontado es extraer el nombre del lugar donde ocurrió el evento. Sin embargo, como se mencionó anteriormente, los resultados que presentamos son preliminares y creemos que podemos mejorar los mismos de la siguiente manera:

1. Experimentando con el tamaño de los contextos, para encontrar el tamaño adecuado para discriminar entre las entidades identificadas.
2. Probar algoritmos de clasificación, para encontrar los que mejor se adapten a la tarea. Para esto hacemos uso del software WEKA [8].

Actualmente, continuamos preparando la colección de contextos para ampliar el componente de extracción de información a los demás eventos naturales propuestos en el dominio de trabajo (ver tabla 1). Como trabajo a futuro pretendemos incorporar a las etapas de preproceso nuevas características, por ejemplo: obtener las partes de la oración. Intentando con esto mejorar los resultados en la tarea de clasificación, y por consiguiente en la extracción.

Finalmente, cabe mencionar que el presente trabajo forma parte del proyecto de investigación *Recolección, Extracción, Búsqueda y Análisis de Información a partir de Textos en Español*, el cual entre sus metas tiene el usar la base de datos sobre desastres en estudios exploratorios y preventivos posteriores.

Agradecimientos

El presente trabajo fue parcialmente financiado por el CONACYT (Proyecto U39957-Y). Asimismo, el primer autor agradece al CONACYT por el apoyo otorgado a través de la Beca para Estudios de Maestría # 171610. Por su parte, el segundo autor agradece a la Secretaría de Estado de Educación y Universidades, España.

Referencias

1. Appelt, D., Israel, D.: Introduction to Information Extraction Technology. A Tutorial Prepared for IJCAI-99 (1999)
2. Chinchor, N.: MUC-7 Test Scores Introduction. In Proceedings of the 7th Message Understanding Conference. Morgan Kaufmann (1997)
3. Grishman, R.: Information Extraction: Techniques and Challenges. Lecture Notes in Artificial Intelligence 1299 (1997)
4. Kushmerick, N., Johnston, E., McGuinness, S.: Information extraction by text classification. Workshop on Adaptive Text Extraction and Mining, Seattle (2001)
5. LA RED: Guía Metodológica de DesInventar. OSSO/ITDG, Lima (2003)
6. Aas, K., Eikvil, L.: Text Categorisation: a Survey. Technical Report, Norwegian Computing Center (1999)
7. Téllez, A., Montes, M., Fuentes, O., Villaseñor, L.: Clasificación Automática de Textos de Desastres Naturales en México. Congreso Internacional en Investigaciones de Ciencias Computacionales, México (2003)
8. Witten, I., Frank, E.: Data Minning: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, Sydney (2000)