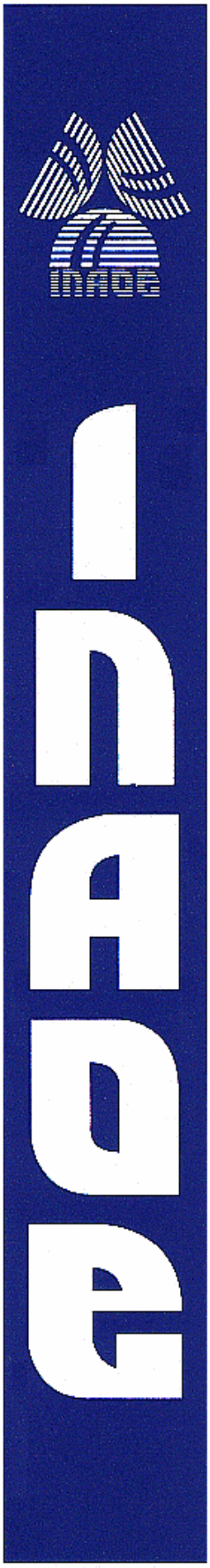

Métodos Basados en Patrones Léxicos para la Extracción de Información

por

Claudia Patricia Orta Palacios

Tesis sometida como requisito parcial
para obtener el grado de

MAESTRA EN CIENCIAS
EN LA ESPECIALIDAD DE
CIENCIAS COMPUTACIONALES

en el
Instituto Nacional de Astrofísica, Óptica y Electrónica.

Supervisada por:

DR. LUIS VILLASEÑOR PINEDA
Coordinación de Ciencias Computacionales, INAOE

DR. MANUEL MONTES Y GÓMEZ
Coordinación de Ciencias Computacionales, INAOE

Tonantzintla, Pue.
2008

© INAOE 2008
Derechos Reservados El autor otorga al INAOE el permiso de reproducir y distribuir copias de esta tesis en su totalidad o en partes

Resumen

Las tecnologías de información actuales han hecho posible el almacenamiento y acceso a grandes colecciones de documentos digitales, pero estas tecnologías aún no han facilitado el análisis de tales cantidades de información. Para satisfacer este requerimiento han surgido recientemente varias tareas de procesamiento de texto. En particular, la extracción de información tiene como fin poblar automáticamente bases de datos mediante la identificación y recolección de piezas de información de documentos de textos libres.

Los trabajos de investigación sobre extracción de información se basan principalmente en el descubrimiento y aplicación de patrones de extracción. Estos trabajos pueden ser clasificados en dos clases principales: métodos supervisados y no-supervisados. El primero hace uso de textos etiquetados en la fase de entrenamiento, mientras que el último evita el uso de tales clases de documentos pero requiere la selección manual y el etiquetamiento de los patrones de extracción descubiertos. En ambos casos es común emplear patrones sintácticos, los cuales crean métodos actuales altamente dependientes del lenguaje.

Este trabajo propone dos diferentes métodos supervisados para la extracción de información. La principal diferencia de dichas propuestas en comparación con métodos previos es que éstas se basan exclusivamente en información léxica y por lo tanto, son fácilmente adaptables a diferentes lenguajes. Además, los métodos propuestos incorporan algunos mecanismos que facilitan la selección y el etiquetamiento manual de los patrones de extracción, haciéndolos muchos más fáciles de mover a diferentes dominios. Los resultados experimentales muestran que el éxito de estos métodos depende del número de patrones léxicos utilizados.

Abstract

Current information technologies have made possible the storage and access to large digital document collections, but they still do not facilitate the analysis of such amounts of information. In order to satisfy this requirement several text processing tasks have recently emerged. In particular, *information extraction* aims to automatically populate databases by identifying and collecting information pieces from free text documents.

The research works on information extraction are mainly based on the discovery and application of extraction patterns. These works can be classified in two main kinds: supervised and not supervised approaches. The formers make use of labeled texts at the training phase, whereas the later ones avoid the use of such kind of documents but require the manually selection and tagging of the discovered extraction patterns. In both cases it is common to employ syntactic patterns, which make current approaches highly language dependent.

This work proposes two different not supervised methods for information extraction. The main difference of these proposals compared with previous approaches is that they are exclusively based on lexical information, and therefore they are easily to adapt to different languages. In addition, the proposed methods incorporate some mechanisms that facilitate the manual selection and tagging of extraction patterns, making them more easily to move to different domains. The experimental results show that the success of these methods depends on the number of used lexical patterns.

Agradecimientos

Un cordial agradecimiento a mis asesores Dr. Luis Villaseñor Pineda y Dr. Manuel Montes y Gómez quienes con su conocimiento, experiencia y sobre todo muy buen humor lograron llevar a buen término este trabajo.

En general se agradece al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) por todas las facilidades prestadas durante mi estancia académica.

Finalmente, se agradece al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado a través de la beca para estudios de maestría no 201829.

Dedicatorias

Para Dios

Porque sin él no hubiera culminado esta etapa de mi vida

Para mi mamá Bernarda.

Por todo su cariño, apoyo y confianza.

"Gracias por confiar en mí mami"

Para Arturo, Celia y Ammy.

Por todo su cariño.

Para Bernardino.

Por todo su cariño, comprensión y sus palabras de ánimo.

Para mis amigos Erika, Gustavo, Javier , Coral y Omar.

Por haberme dado su confianza y haberme apoyado en los momentos difíciles.

"Mil gracias"

Contenido

Capítulo 1

Introducción

La era tecnológica en la que vivimos ha facilitado la acumulación y acceso a grandes colecciones de documentos digitales. Sin embargo, el análisis manual de la información contenida en estas colecciones es en extremo difícil. De ahí la necesidad de buscar medios automáticos.

Desde hace muchos años el tratamiento automático de textos ha sido un área activa de investigación. Dentro de esta área se han desarrollado diversos campos para tratar sus diversas problemáticas. Algunos de estos campos son: la *Recuperación de Información*, la *Búsqueda de Respuestas,* la *Minería de Texto*, la *Clasificación de Textos*, la *Extracción de Información*, entre otros.

El presente trabajo cae dentro del campo de la *Extracción de Información.* Ésta tiene como fin identificar y recolectar piezas de información inmersas en los textos con el objetivo de llenar un conjunto de plantillas. Cabe mencionar, que una plantilla se define como un gran marco de caso compuesto por un conjunto de ranuras predefinidas para cada pieza de información que debe

ser extraída del texto. La idea de llenar plantillas es facilitar el acceso y análisis de grandes cantidades de datos por medios automáticos.

Como es de imaginar la tarea de “extraer” información es complicada. Básicamente la tarea recae en la identificación de los datos a extraer gracias a un conjunto de *patrones de extracción*. Dichos patrones cubren las distintas formas lingüísticas en que las piezas de información se expresan en los textos. Desafortunadamente, la generación de dichos patrones aún no ha sido del todo resuelta. Los primeros trabajos intentaron determinar de forma manual los patrones de extracción, para ello uno o varios expertos distinguían los patrones pertinentes después de revisar un número “suficiente” de documentos. El resultado era un conjunto de patrones sintácticos para cada campo a extraer. Por supuesto, la tarea era ardua y dependiente tanto del dominio como del tipo de información a extraer. Decidir la extracción de una nueva pieza de información involucraba una nueva revisión para determinar el nuevo conjunto de patrones. Desafortunadamente, aún después de este costoso proceso no era posible asegurar que el conjunto de patrones de extracción fuera completo ni tampoco asegurar la precisión de dichos patrones.

Trabajos subsecuentes se enfocaron en el descubrimiento automático de patrones. Estos los podemos agrupar en dos grandes líneas: el enfoque supervisado y el enfoque no-supervisado. El primer enfoque parte de corpus previamente etiquetados. Es decir, corpus donde las piezas de información a extraer han sido indicadas. Posteriormente, un proceso automático determina posibles patrones. De este conjunto de patrones candidatos se seleccionan manualmente los patrones apropiados. Así se reduce el trabajo del experto, al no realizar la revisión manual de cientos de documentos, concentrando sus esfuerzos en depurar el conjunto propuesto de patrones candidatos. El segundo enfoque evita el uso de corpus etiquetados. Para ello, el proceso

automático incluye un paso para determinar automáticamente las posibles piezas de información relevante. Posteriormente, al igual que el enfoque supervisado, un proceso manual selecciona los patrones. En este caso, la selección de los patrones tiene una doble finalidad: determinar si un patrón es apropiado y determinar el tipo de información que dicho patrón extrae.

El presente trabajo propone dos métodos no-supervisados de extracción de información. Estos métodos intentan disminuir su dependencia al dominio e inclusive al idioma de los documentos. En la siguiente sección se introduce y acota el problema abordado por esta tesis. Después, en la sección 1.2 se describe brevemente la solución propuesta al problema planteado. Enseguida, en la sección 1.3 se exponen los objetivos de la tesis, por último en la sección 1.4 se describe brevemente la organización de la tesis.

1.1 Descripción del problema

El problema principal de la Extracción de Información es el descubrimiento de los patrones de extracción. Como se mencionó existen dos grandes enfoques: el enfoque supervisado y el enfoque no-supervisado. En el enfoque no supervisado, el cual es en el que se basa esta tesis, surge el problema de la expresividad de los patrones descubiertos. La gran mayoría, sino la totalidad, de los métodos no-supervisados descubren patrones sintácticos. Esto les da a los patrones un gran poder de expresividad, es decir, un patrón es capaz de capturar un gran número de formas lingüísticas en que la información puede aparecer. Sin embargo, al depender de información sintáctica el proceso de descubrimiento de patrones se hace dependiente del rendimiento de las herramientas lingüísticas a nuestro alcance. Por otro lado, el proceso de selección manual se complica pues los

patrones no son legibles para cualquier persona, por lo que es necesario contar con un experto en el dominio con sólidas nociones lingüísticas.

1.2 Solución propuesta

Este trabajo presenta dos métodos de extracción de información basados en un enfoque no-supervisado. El objetivo de ambos métodos es facilitar su portabilidad a otros dominios. Para ello, se orientan al descubrimiento de patrones de extracción a nivel léxico. Con esto se disminuye la dependencia de ambos métodos a herramientas lingüísticas y, por otro lado, simplificarán las exigencias impuestas al experto durante el proceso de selección de patrones.

Finalmente, debido a la utilización de patrones léxicos, dichos métodos serán más fáciles de trasladar a otros dominios e inclusive a otros lenguajes.

1.3 Objetivos

Objetivo General

Proponer dos métodos no-supervisados de extracción de información utilizando patrones léxicos.

Objetivos específicos

- Desarrollar un método para descubrir patrones léxicos de extracción.
- Desarrollar un método para facilitar la categorización manual de los patrones léxicos de extracción.

- Diseñar un método para el llenado de plantillas aplicando los patrones léxicos de extracción.
- Evaluar los métodos de extracción de información propuestos en dos dominios diferentes.

1.4 Organización de la tesis

En el capítulo 2 se presentan los conceptos básicos necesarios para la comprensión de los métodos propuestos. En el capítulo 3 se presenta el estado del arte en la tarea de Extracción de Información. Así como la descripción de los métodos de evaluación empleados para calificar los sistemas de extracción de información. El capítulo 4 presenta un primer método no-supervisado basado en patrones de extracción de tamaño variable mostrando los resultados alcanzados. Mientras el capítulo 5 describe un segundo método basado en patrones de extracción de tamaño fijo y presenta sus resultados. Finalmente, en el capítulo 6 se presentan las conclusiones y el trabajo futuro que se desprende de esta tesis.

Capítulo 2

Conceptos básicos

En este capítulo se introducen los conceptos básicos necesarios para entender los siguientes capítulos. En la sección 2.1, se muestra la definición de la técnica empleada para calcular las Secuencias Frecuentes Maximales útiles para el descubrimiento de patrones léxicos, así como una explicación del funcionamiento del algoritmo de agrupamiento estrella. Este algoritmo es utilizado en esta tesis para encontrar aquellos patrones léxicos descubiertos útiles para facilitar la tarea de etiquetado manual de los patrones.

2.1 Minería de texto

La minería de texto puede definirse como un proceso de conocimiento-intensivo, en el cual un usuario interactúa con una colección de documentos por encima del tiempo, mediante el uso de un conjunto de herramientas de análisis [Feldman R. & Sanger J., 2007].

De manera análoga a la minería de datos, la minería de textos pretende extraer información útil de fuentes de datos, mediante la identificación y exploración de patrones interesantes. En el caso de la minería de texto, sin

embargo, las fuentes de datos son colecciones de documentos y los patrones de interés no son encontrados entre registros de bases de datos formalizadas, sino en los datos textuales no estructurados de los documentos de estas colecciones.

Ciertamente, la minería de texto deriva mucha de su inspiración y dirección de la investigación seminal sobre minería de datos [Feldman R. & Sanger J., 2007]. Por lo tanto, no es de sorprender que se hagan evidentes muchas similitudes de alto nivel en cuestión de arquitectura entre los sistemas de minería de texto y los de minería de datos. Por ejemplo, ambos tipos de sistemas se basan en rutinas de pre-procesamiento, algoritmos de descubrimiento de patrones y elementos de representación de capas.

En la minería de datos y de texto se emplea frecuentemente la tarea denominada agrupamiento [Hernández J. et al., 2004]. Esta tarea sirve para dividir una colección de objetos no etiquetados dados dentro de grupos sin ninguna información anticipada, de tal forma que los objetos dentro de un grupo tengan una similaridad alta y que además, éstos sean muy distintos a los objetos de los otros grupos [Han J. & Kamber M.,2001]. Es importante señalar que la tarea de agrupamiento sólo es apropiada cuando se dé la mínima información acerca de los datos y la decisión del autor se deba realizar con el mínimo de suposiciones posibles acerca de éstos. En esta tesis se empleó el agrupamiento para formar grupos de oraciones sin etiquetar. En la sección 2.1.2, se describe más a detalle los algoritmos de agrupamiento existentes, así como el algoritmo utilizado en esta tesis.

Por otro lado, la minería de texto consta de un núcleo de operaciones para el descubrimiento de conocimiento, el cual se centra en algoritmos útiles para descubrir patrones en colecciones de documentos. Dicho *núcleo de operaciones para el descubrimiento del conocimiento* consta de varios

mecanismos. Ejemplos de dichos mecanismos son: Distribuciones, Asociaciones y Secuencias Frecuentes Maximales (SFM). Dichas SFM pueden ser usadas como elementos básicos para el descubrimiento de conocimiento [Ahonen-Myka H., 1999]. En esta tesis se utilizaron las SFM para descubrir patrones léxicos de extracción.

2.1.1 Secuencias Frecuentes Maximales (SFM)

Una secuencia frecuente maximal es una secuencia de palabras que debe aparecer en un número dado (umbral) de ejemplos (por ejemplo, documentos, oraciones, etc.) y además, no debe estar contenida en otra secuencia de palabras.

A continuación, se presenta la definición formal de Secuencias Frecuentes Maximales. Para entender ésta es necesario asumir que *D* es un conjunto de textos (por texto nos referimos a un documento completo o incluso a una sola oración), donde cada texto consiste de una secuencia de palabras [Ahonen-Myka H., 2002].

Definición 1. Una secuencia $p = a_1 \ldots a_k$ es una subsecuencia de una secuencia q si todos los elementos a_i, $1 \leq i \leq k$, ocurren en q y además, ocurren en el mismo orden que en p. Si una secuencia p es una subsecuencia de una secuencia q, entonces se dice que p ocurre en q.

Definición 2. Una secuencia p es *frecuente* en D si p es una subsecuencia de al menos σ textos de D, donde σ es un umbral de frecuencia predefinido.

Definición 3. Una secuencia *p* es una *secuencia frecuente maximal* en *D* si no existe alguna otra secuencia *p´* en *D* tal que *p* es una subsecuencia de *p´* y *p´* es frecuente en *D*.

Para ejemplificar, considérese el conjunto de oraciones extraídas de un corpus de desastres naturales (véase la figura 2.1).

1. MILES DE MUERTOS SE ENCONTRARON EN LA REGION AFGANA
2. EN LA REGION DEL CAIRO MURIERON MUCHAS PERSONAS A RAIZ DEL TEMBLOR
3. EN MEXICO MILES DE MUERTOS QUEDARON POR LAS CALLES
4. EN TOTAL MURIERON MUCHAS PERSONAS POR EL INCENDIO
5. SE DERRUMBARON MILLONES DE CASAS POR EL SISMO
6. EN LAS CERRANIAS SE DERRUMBARON MILLONES DE CASAS
7. MILES DE HOGARES FUERON AFECTADOS EN ARGENTINA
8. MUCHAS PERSONAS QUEDARON INCOMUNICADAS
9. AUNQUE MUCHAS PERSONAS FUERON DAMNIFICADAS
10. EN COZUMEL MILES DE HOGARES FUERON AFECTADOS POR LA INUNDACION

Figura 2.1. Conjunto de oraciones referentes al tema “desastres naturales”.

Si se sacarán las secuencias frecuentes maximales de estas oraciones con un umbral de 2, el resultado sería el que se muestra en la figura 2.2.

2 SFM de tamaño 1
    [2] QUEDARON
    [2] LAS
1 SFM de tamaño 2
    [2] POR EL
3 SFM de tamaño 3
    [2] MILES DE MUERTOS
    [2] EN LA REGION
    [2] MURIERON MUCHAS PERSONAS
2 SFM de tamaño 5
    [2] MILES DE HOGARES FUERON AFECTADOS
    [2] SE DERRUMBARON MILLONES DE CASAS

Figura 2.2. SFM obtenidas.

De acuerdo con la figura 2.2, se obtuvieron un total de 8 SFM. De dichas SFM, dos son de tamaño uno, una es de tamaño dos, tres son de tamaño tres y dos son de tamaño cinco.

Como se mencionó anteriormente, una secuencia frecuente maximal es aquella que debe aparecer en un número dado (umbral) de oraciones y además, no debe estar contenida en otra secuencia de palabras. Para este ejemplo se empleo el valor de dos como umbral. Por lo tanto, el primer paso fue buscar aquellas secuencias de oraciones que aparecieran como mínimo en dos oraciones de las diez de entrada (también denominadas secuencias frecuentes).

Una vez que ya se tenían las secuencias frecuentes, el siguiente paso era identificar cuáles de éstas eran maximales. Es decir, que no estuvieran contenidas en otras secuencias de palabras. Es por ello que la secuencia “MUCHAS PERSONAS” no aparece como SFM. Esto se debe a que dicha secuencia aunque cumple con aparecer como mínimo dos veces en las oraciones de entrada, sin embargo, está contenida dentro de la secuencia “MURIERON MUCHAS PERSONAS”.

Según [Kovács L. & Ahonen-Myka H.,2001] la razón para extraer SFM en lugar de secuencias de tamaño fijo es porque las SFM tienen una representación flexible y compacta.

El algoritmo empleado para extraer Secuencias Frecuentes Maximales (DIMASP) en este trabajo fue desarrollado por René A. García Hernández (para un estudio más completo véase [García R. et al.,2006]).

2.1.2 Algoritmos de agrupamiento

Los algoritmos de agrupamiento sirven para dividir una colección de objetos no etiquetados en grupos sin ninguna información anticipada. En esta tesis,

se agruparon oraciones sin etiquetar. Por tal motivo, fue necesario elegir un algoritmo de agrupamiento.

Existen varios algoritmos de agrupamiento en la actualidad. Sin embargo, la elección del adecuado depende de dos aspectos: los datos disponibles y la aplicación.

Los principales métodos de agrupamiento se dividen en dos grupos: *Jerárquicos y Particionales* [Jain A.et al.,1999]. A continuación se describen cada uno de éstos.

- Los *algoritmos jerárquicos* crean una descomposición jerárquica del conjunto de objetos de datos dado. Este tipo de agrupamiento también se puede dividir en *agrupamiento jerárquico aglomerativo y agrupamiento jerárquico divisivo,* dependiendo de si la descomposición jerárquica se forma de abajo hacia arriba o de arriba hacia abajo.

    - *Agrupamiento jerárquico aglomerativo*. Esta estrategia que va de abajo hacia arriba inicia estableciendo a cada objeto como un grupo. Posteriormente, se mezclan los dos grupos atómicos más similares aun disponibles. Este procedimiento continúa hasta que todos los objetos estén dentro de un grupo simple o hasta que ciertas condiciones de término se satisfagan.

    - *Agrupamiento jerárquico divisivo*. Esta estrategia de arriba hacia abajo hace lo inverso al agrupamiento jerárquico aglomerativo, ya que éste inicia con todos los objetos en un grupo. Este método subdivide el grupo en piezas más y más pequeñas, hasta que cada objeto forme un grupo sobre sí mismo o hasta que se satisfagan ciertas condiciones de término, como por ejemplo, que ya se hayan

obtenido el número deseado de grupos o que la distancia entre los dos grupos más cercanos este por encima de un cierto umbral.

Dentro de los *algoritmos jerárquicos* existentes se pueden mencionar los siguientes: BIRCH, CURE, Chameleon, COBWEB, entre otros.

- Los *algoritmos particionales* funcionan de la siguiente manera, dada una base de datos de *n* objetos o tuplas de datos, dichos algoritmos construyen *k* particiones de los datos, donde cada una de éstas representa un grupo y $k \leq n$, *d*e tal forma que los datos clasificados dentro de los *k* grupos satisfacen los siguientes requerimientos: (1) cada grupo debe contener al menos un objeto y (2) cada objeto debe pertenecer a exactamente un grupo [Han J. & Kamber M.,2001].

  Los *algoritmos particionales* se dividen en dos grupos: los que necesitan que se les indique la cantidad de grupos a formar (k) y los que no requieren que se les indique el número de grupos a formar. Es importante resaltar que los algoritmos del último grupo inducen de manera natural el número de grupos a formar. Algunos ejemplos de *algoritmos particionales* que se ubican en el primer grupo son: k-Means y k-Medoids. Para el caso de los *algoritmos particionales* que no requieren una k de entrada, se puede mencionar como ejemplo el *algoritmo estrella*.

  El *algoritmo estrella* se caracteriza por inducir de manera natural el número de grupos a formar (k). De acuerdo con [Aslam J. et. al.,2000] este algoritmo es altamente eficiente y simple de implementar. Y además, garantiza la calidad de los grupos y calcula más grupos exactos [Aslam J. et al., 1999].

En esta tesis, se utilizó el algoritmo estrella desarrollado por [Aslam J. et al., 1999]. Se empleó dicho algoritmo principalmente porque no requiere como entrada el número de grupos a formar y porque indica de manera automática el objeto más representativo de cada grupo.

2.1.2.1 Algoritmo estrella

El algoritmo de agrupamiento estrella (véase la versión resumida de este algoritmo en la figura 2.3) se basa en una cobertura del grafo de similaridad umbralizado G_σ por medio de subgrafos en forma de estrella. Entiendase G_σ como aquel grafo no dirigido obtenido de G donde éste es un grafo no dirigido ponderado $G = (V, E, w)$. En G los vértices del grafo corresponden a los textos y cada arista ponderada corresponde a la similitud que existe entre dos documentos. Cabe mencionar que existen varias medidas para calcular la similitud entre dos documentos, siendo las más conocidas las siguientes:

- *Medida Cosenoidal*. La idea básica de esta medida consiste en medir el ángulo entre el vector de D_i y de D_j. Para hacer esto se calcula lo siguiente:

$$SC(D_i, D_j) = \frac{\sum_{k=1}^{t} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{t} (w_{jk})^2 \sum_{k=1}^{t} (w_{ik})^2}} \tag{2.1}$$

- *Medida Dice*. El coeficiente de Dice se obtiene mediante la siguiente ecuación:

$$SC(D_i, D_j) = \frac{2\sum_{k=1}^{t} w_{ik} w_{jk}}{\sum_{k=1}^{t} (w_{jk})^2 + \sum_{k=1}^{t} (w_{ik})^2} \tag{2.2}$$

- *Medida de Jaccard*. El coeficiente de Jaccard se calcula empleando la siguiente ecuación:

$$SC(D_i, D_j) = \frac{\sum_{k=1}^{t} w_{ik} w_{jk}}{\sum_{k=1}^{t} (w_{jk})^2 + \sum_{k=1}^{t} (w_{ik})^2 - \sum_{k=1}^{t} w_{ik} w_{jk}} \quad (2.3)$$

Donde en cada uno de los casos (véanse ecuaciones 2.1, 2.2 y 2.3):

k : va de uno al número total de términos del vocabulario t.

w_{ik} : es el peso del término k en el documento D_i, o 0 (cero) si el documento D_i no tiene el término.

w_{jk} : es el peso del término k en el documento D_j, o 0 (cero) si el documento D_j no tiene el término.

Para cualquier umbral σ:

1. Calcular $G_\sigma = (V, E_\sigma)$ donde $E_\sigma = \{e \in E : w(e) \geq \sigma\}$
2. Poner cada vértice en G_σ inicialmente como *no marcado*.
3. Calcular el grado de cada vértice (o número de aristas) $v \in V$.
4. Tomar el vértice de mayor grado que tenga la etiqueta "no-marcado" como centro de la estrella y sus vértices asociados como satélites. Marcar cada nodo de la estrella recién construida.
5. Repetir el paso 4 hasta que todos los vértices estén marcados.
6. Representar cada grupo por medio del documento correspondiente al centro de cada estrella.

Figura 2.3. Algoritmo estrella.

Capítulo 3

Estado del arte

El presente capítulo tiene como objetivo presentar al lector en primer lugar el área en la que se ubica esta tesis, así como algunos de los trabajos que se han llevado a cabo en dicha área. De igual forma, se presenta una descripción de las características generales con que cuentan dichos trabajos, los cuales se caracterizan por emplear dos tipos de enfoques: el supervisado y el no-supervisado, logrando con ello informar al lector del porque es importante el desarrollo de los métodos de extracción de información presentados en esta tesis. De igual manera, se describe la forma en que se han evaluado los sistemas de extracción de información existentes.

3.1 Extracción de información

3.1.1 Definición

La extracción de información es la tarea que se encarga de identificar descripciones de eventos en textos en lenguaje natural y por consiguiente, extraer la información relacionada a dichos eventos [Patward S. & Riloff E.,2006]. En otras palabras, un sistema de extracción de información

encuentra y enlaza la información relevante, mientras ignora la extraña e irrelevante [Cowie J. & Lehnert W., 1996].

Los inicios de la extracción de información se ubican a mediados de los años 60´s. Sin embargo, es a finales de los 80´s cuando esta tecnología comienza a tener auge. Esto se debió a tres factores: el poder computacional, el exceso de información textual existente de forma electrónica y la intervención de la Agencia de Defensa de los Estados Unidos (DARPA).

> DARPA patrocinó durante los años de 1987 a 1998 las siete conferencias sobre entendimiento de mensajes (MUC). Asimismo, durante los años de 1990 a 1998 DARPA activó el TIPSTER (Programa de Investigación sobre Recuperación y Extracción de Información), donde las MUC fueron incluidas.

Las MUC fueron las que inicialmente fomentaron las competencias entre distintos grupos de investigación. Las cuales se llevaron a cabo con el objetivo de desarrollar sistemas de extracción de información. Es por ello que también definieron sus propios métodos de evaluación. En cada una de las MUC se han empleado diferentes dominios. En MUC-1 y MUC-2 se utilizaron noticias sobre operaciones navales, posteriormente, en MUC-3 y MUC-4 se empleó el dominio sobre atentados terroristas en América Latina. Después, en MUC-5 [Chinchor N. & Sundheim B., 1993] se hizo uso de noticias sobre fusiones de empresas y anuncios de productos microelectrónicos. De igual forma, en MUC-6 [Sundheim B., 1993] se utilizaron noticias sobre sucesión de directivos. Asimismo, en MUC-7 [Chinchor N., 1998] hicieron uso de dos dominios, uno sobre noticias de accidentes de avión y otro sobre lanzamiento de misiles y artefactos (para un estudio más completo véase [Grishman R., 1993]).

A continuación, se muestra un ejemplo de cómo sería el funcionamiento de un sistema de extracción de información. El siguiente texto es una parte de un documento que pertenece al dominio de sucesión de directivos extraído de un texto libre [Turmo J. et al., 2006].

> *A.C.Nielsen Co. dijo que George Garrick, de 40 años, presidente de los recursos de información de Londres que se basa en la operación de servicios de información europea, se convertirá en presidente de Nielsen Marketing Research, una unidad de la corporación Dun&Bradstreet. Él será el sucesor de John I.Costello quién renunció en marzo.*

La salida de un sistema de extracción de información es un conjunto de registros por noticia de entrada. En la tabla 3.1, se muestra el registro extraído del fragmento de texto mostrado en esta sección. Cabe mencionar que cada registro está compuesto por campos. Dichos campos se establecen desde las primeras etapas del sistema de extracción y se agrupan en lo que se denomina plantilla de extracción. Es importante señalar que cada campo representa información relevante de acuerdo al dominio, la cual será útil para el análisis del conjunto de documentos textuales de entrada.

INFORMACIÓN DE DIRECTIVOS	
PERSONA ENTRANTE	*George Garrick*
PERSONA SALIENTE	*John I.Costello*
PUESTO	*Presidente*
ORGANIZACIÓN	*Nielsen Marketing Research*

Tabla 3.1. Registro generado por un sistema de extracción de información.

Se han construido diversos métodos de extracción de información hasta la fecha. No obstante, los trabajos que presentan dichos métodos se caracterizan por emplear dos tipos de enfoques: el supervisado y el no-supervisado.

3.2 Enfoque supervisado

Los métodos de extracción de información basados en el enfoque supervisado se caracterizan por recibir como entrada un conjunto de documentos etiquetados. A partir de dichos textos se descubren un conjunto de patrones de extracción para un dominio específico. Estos patrones sirven para extraer información de textos pertenecientes al mismo dominio.

Este tipo de métodos constan de cuatro fases para su construcción. Dichas fases son las siguientes: *Etiquetado, Pre-procesamiento, Descubrimiento de Patrones de Extracción y Generación de Plantillas.* En la figura 3.1, se muestra la arquitectura de este tipo de métodos.

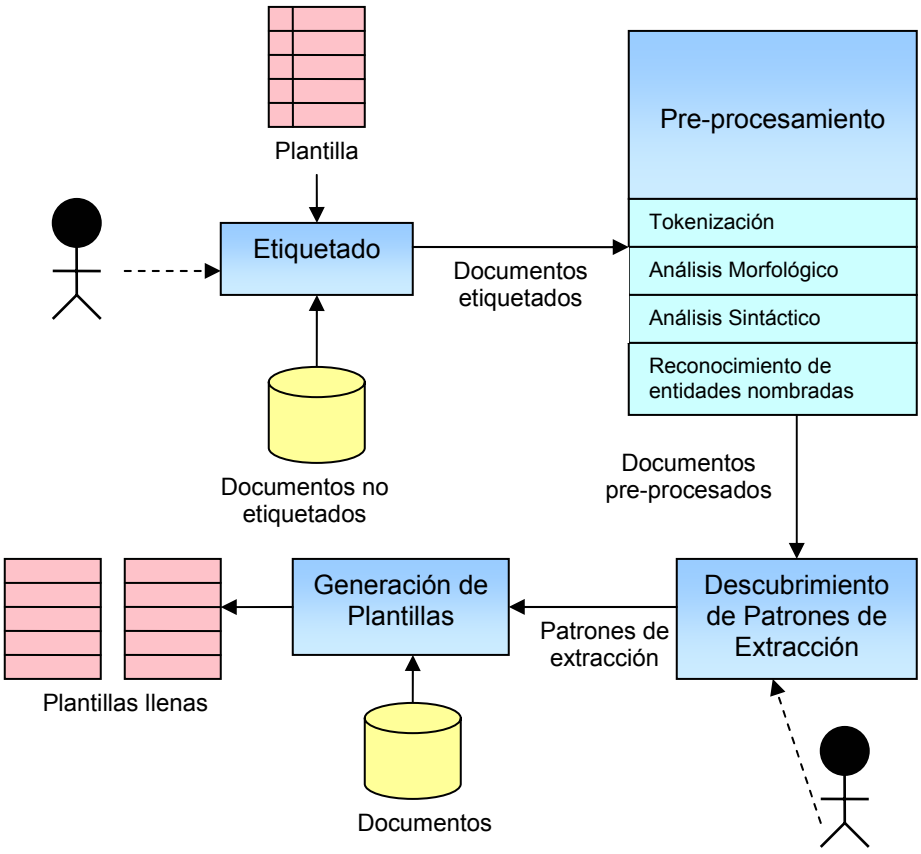

Figura 3.1. Arquitectura de los métodos de extracción de información basados en el enfoque supervisado.

De acuerdo con la figura 3.1, los métodos de extracción de información construidos bajo este enfoque necesitan de una fase de *Etiquetado* de documentos. Esta tarea se realiza en forma manual con base en la plantilla dada. De acuerdo con [Riloff E., 1996] dicha tarea se lleva a cabo en un tiempo aproximado de ocho horas para etiquetar solamente 160 documentos. Por lo tanto, el tiempo de construcción de este tipo de sistemas esta muy relacionado con la cantidad de documentos que el sistema reciba como entrada.

La fase *Pre-procesamiento* que se muestra en la figura 3.1, tiene como objetivo identificar fragmentos de textos de los documentos de entrada. Esta tarea se realiza ya que es necesario identificar en los textos de entrada los elementos que se pretenden extraer. Los fragmentos de texto resultantes serán útiles para descubrir patrones de extracción a partir de éstos.

Durante la fase *Pre-procesamiento* se efectúan principalmente algunas de las cuatro tareas siguientes: Tokenización, Análisis Morfológico, Análisis Sintáctico y Reconocimiento de Entidades Nombradas. A continuación se describen las actividades que realizan cada una de éstas.

Tokenización. Divide un documento de entrada en bloques de construcción básica (palabras, sentencias y párrafos).

Análisis Morfológico. Su función consiste en detectar la relación que se establece entre las unidades mínimas que forman una palabra, como puede ser el reconocimiento de sufijos o prefijos. Este nivel de análisis mantiene una estrecha relación con el análisis léxico.

Análisis Sintáctico. Tiene como función etiquetar cada uno de los componentes sintácticos que aparecen en la oración y analizar cómo las

palabras se combinan para formar construcciones gramaticalmente correctas. El resultado de este proceso consiste en generar la estructura correspondiente a las categorías sintácticas formadas por cada una de las unidades léxicas que aparecen en la oración.

Reconocimiento de entidades nombradas. Identifica en los documentos elementos como por ejemplo, nombres de personas, lugares y organizaciones.

Asimismo, siguiendo con la arquitectura mostrada en la figura 3.1. La fase *Descubrimiento de Patrones de Extracción* tiene como objetivo generar un conjunto de patrones de extracción para su utilización en la tarea de extracción de información. Para poder lograr el objetivo, es necesario realizar un proceso de correspondencia de tipos con respecto a la plantilla dada. Este proceso se realiza manualmente por un experto y su fin es asignar el tipo de información que extraerá cada patrón descubierto.

Finalmente, la fase *Generación de Plantillas* tiene como objetivo generar una plantilla llena por cada documento de entrada, para lograrlo se lleva a cabo la tarea de extracción de información en los textos dados, utilizando los patrones de extracción descubiertos en la fase anterior.

Entre los trabajos más representativos que se han realizado con respecto a este tema están los que se describen a continuación.

En primer lugar, está el trabajo desarrollado por [Riloff E., 1993]. Dicho trabajo consistió en extraer información con base en la técnica denominada “extracción de conceptos selectivos”, implementada en un sistema llamado AutoSlog. Este sistema construía automáticamente un diccionario de nodos concepto para el dominio de “atentados terroristas”.

De acuerdo con la arquitectura mostrada en la figura 3.1, en el trabajo de Ellen Riloff, se realizó la fase de *Etiquetado* de un conjunto de 1500 textos sobre atentados terroristas en América Latina. También, se efectuó la fase de *Pre-procesamiento.* Durante dicha fase se hicieron las tareas de Tokenización y Análisis Sintáctico. Una vez pre-procesados los textos, durante la fase de *Descubrimiento de Patrones de Extracción* se identificaron en dichos textos patrones lingüísticos. Un ejemplo de un patrón lingüístico utilizado en este trabajo es el siguiente: ***<sujeto> verbo en pasivo***. Para poder lograr lo anterior, usaron 13 heurísticas construidas manualmente. El resultado fueron palabras de activación para los nodos concepto. Posteriormente, un experto revisó 1237 definiciones de nodos concepto, con el fin de seleccionar sólo las útiles. El resultado fueron 450 definiciones de nodos concepto. Los nodos concepto resultantes funcionaban como patrones de extracción durante la fase de *Generación de Plantillas*. En la tabla 3.2, se muestra un ejemplo de un nodo concepto para extraer el responsable del atentado. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión, recuerdo y medida-F (en la sección 3.4, se explican dichas métricas).

NODO CONCEPTO	
Nombre:	Target-subject-passive-verb-bombed
Palabra de activación:	Bombed
Ranuras variables:	(target(*S*1))
Restricciones:	(class phys-target *S*)
Ranuras constantes:	(type bombing)
Condiciones permitidas:	((passive))

Tabla 3.2. Definición de un nodo concepto.

Asimismo, en [Soderland S. et. al., 1995] se presentó *CRYSTAL,* el cual es un sistema que aprende reglas de extracción de un conjunto de textos de entrenamiento.

De acuerdo con las fases explicadas al inicio, en dicho trabajo se llevó a cabo una fase de *Etiquetado* de 385 documentos sobre pacientes dados de alta en hospitales. De igual manera, se efectuó una fase de P*re-procesamiento* de los documentos de entrada. En dicha fase se hicieron las tareas de Tokenización, Análisis Morfológico y Sintáctico. Posteriormente, para la fase de *Descubrimiento de Patrones* de Extracción se usó un algoritmo de cobertura para aprender reglas de extracción y sus complicados elementos restrictivos. El resultado fue un conjunto de nodos concepto específicos para cada frase que se extrajera en el corpus de entrenamiento, es por esto, que se requería de conocimiento de fondo en forma de jerarquía de 133 clases específicas al dominio, así como de un léxico con información de las clases semánticas para 95,000 términos. Al igual que como se muestra en la arquitectura de la figura 3.1, los nodos concepto resultantes en este trabajo funcionaron como patrones de extracción durante la última fase de dicha arquitectura. En la figura 3.2, se muestra un ejemplo de una definición de nodo concepto utilizado en este trabajo para la reaparición de una enfermedad. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión y recuerdo.

Tipo de Nodo Concepto: Diagnosis
Subtipo: Pre-existencia
Extraer de frase preposicional. —WITH”
Verbo en voz pasiva
Restricciones del verbo:
Incluir palabras—DIAGNOSED”
Restricciones de frase preposicional:
Preposición =— WITH
Incluir palabras—RECURRENCE OF”
Modificador de clase <Body part or Organ>
Cabecera de la clase <disease or Syndrome>

Figura 3.2. Definición de nodo concepto para la reaparición de una enfermedad.

Posteriormente, en la investigación realizada por [Ciravegna F., 2001] se propuso el algoritmo de capa para la extracción de información adaptable de textos denominado $(LP)^2$.

De acuerdo con la arquitectura mostrada en la figura 3.1, en el trabajo de Fabio Ciravegna se realizó una fase de *Etiquetado*. Esta fase se efectuó debido a que el algoritmo propuesto requería de ejemplos positivos y negativos. Por tal motivo, se identificaron ejemplos positivos de un corpus de entrenamiento compuesto por 485 documentos. Los ejemplos positivos eran las etiquetas SGML (Lenguaje de Marcación Generalizado) (como por ejemplo, <speaker>) insertadas por el experto, de tal forma que el resto del corpus de entrenamiento se consideró una piscina de ejemplos negativos. Posteriormente, se efectúo una fase de *Pre-procesamiento*. En dicha fase se hicieron las tareas de Tokenización, Análisis morfológico y Reconocimiento de Entidades Nombradas, de tal forma que por cada ejemplo positivo, el algoritmo primero construía una regla inicial y después generalizaba la regla. Asimismo, se realizó una fase de *Descubrimiento de Patrones de Extracción*, la cual consistía en mantener las k mejores generalizaciones de la regla inicial. Éstas servirían como patrones de extracción. Un ejemplo de una regla para extraer la hora se muestra en la tabla 3.3. Finalmente, para la fase de *Generación de Plantillas*, el algoritmo $(LP)^2$ se implementó en el sistema denominado *LearningPinocchio* para extraer información de resúmenes profesionales escritos en inglés. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión, recuerdo y medida-F.

Índice de palabras	**Condición**					**Acción**
	palabra	lema	lexcat	caso	semcat	Etiqueta
3		at				**<time>**
4			Dígito			
5					TiempoId	

Tabla 3.3. Regla generalizada.

De igual forma, en [Callan J. & Mitamura T.,2002] se presentó KENE, un sistema que tomaba un método de generación y prueba para la extracción de entidades nombradas de documentos. En dicho trabajo se efectuaron de igual manera las cuatro fases de la arquitectura mostrada en la figura 3.1. Durante la fase de *Etiquetado* se identificaron 93,989 nombres de autores y 9292 de organizaciones. El corpus empleado estaba formado de citas de publicaciones científicas en Ciencias de la Computación y de páginas de autores identificadas en el sitio Web oficial de la Universidad de Trier [Hoff G., 2002]. Después, se realizó la fase de *Pre-procesamiento*, en la que cada documento era analizado gramaticalmente. Durante esta fase se realizaron las tareas de Tokenización y Análisis Sintáctico. Asimismo, se efectuó la fase de *Descubrimiento de Patrones de Extracción*. Durante dicha fase se produjo un conjunto de reglas de extracción candidatas como resultado del análisis gramatical realizado. Es importante señalar que cada regla de extracción era un camino de un árbol de análisis sintáctico (secuencia de etiquetas marcadas en el documento y signos de puntuación) que alcanzaba alguna cadena. Enseguida, se aplicaron las reglas de extracción candidatas sobre los textos de entrada, para posteriormente, buscar en una base de datos de entidades nombradas las cadenas extraídas por cada regla, de tal forma que aquella regla que extrajera un número específico de nombres conocidos se consideraba válida y se agregaba en una base de datos. Finalmente, durante la fase de *Generación de Plantillas* las reglas resultantes podían utilizarse para extraer entidades nombradas de los textos de entrada. Un ejemplo de una regla utilizada en este trabajo se muestra en la figura 3.3. Esta regla especifica un camino a través de un árbol sintáctico que finaliza con una lista separada por comas. Esta regla extrae el segundo elemento de la lista. Es importante señalar que las palabras body, table, tr, td, p y br son etiquetas HTML. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión, recuerdo y medida-F.

ID de regla:	62
Padre:	22
Sintaxis:	body/table/tr/td/p/br/list[1]-comma
Aplicaciones:	39 veces

Figura 3.3. Regla de extracción.

Otro de los trabajos realizados con respecto a este tema es el de [Califf M. & Mooney R., 2003]. En este trabajo se presentó el algoritmo RAPIER. Dicho algoritmo empleaba pares de documentos prueba y realizaba el llenado de plantillas. Esto último con el propósito de generar reglas útiles para la tarea de extracción de información. En dicho trabajo se llevaron a cabo las fases explicadas al inicio de esta sección. Durante la fase de *Etiquetado* se realizó un llenado manual de los registros de una base de datos de textos, ya que el algoritmo RAPIER necesitaba de éstos para su funcionamiento. Después, durante la fase de *Pre-procesamiento* se efectuó un Análisis Morfológico y Sintáctico. Asimismo, durante la fase de *Descubrimiento de Patrones de Extracción* se empleó el aprendizaje de abajo hacia arriba para construir reglas de extracción útiles. La representación de las reglas de extracción RAPIER son similares a Eliza [Weizenbaum J.,1966]. Cada regla de extracción está formada de tres patrones: (1) un patrón pre-relleno que debe corresponder con el texto inmediatamente después del rellenado, (2) un patrón de relleno que debe coincidir con la actual ranura de relleno y (3) un patrón de post-relleno que debe coincidir con el texto inmediatamente después del rellenado. Un ejemplo de un patrón utilizado en este trabajo se muestra en la tabla 3.4, donde nn y nnp son las partes de las etiquetas del discurso para sustantivo y sustantivo apropiado. En cambio, jj es parte de las etiquetas del discurso para un adjetivo. Finalmente, las reglas de extracción generadas se utilizaron durante la fase de *Generación de Plantillas* sobre dos dominios (trabajos por computadora y seminarios). El corpus sobre trabajos de computadora estaba integrado por 300 documentos, en cambio el

segundo corpus estaba compuesto por 485 documentos. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión y recuerdo.

Patrón de pre-relleno	Patrón de relleno	Patrón de post-relleno
1) sintáctico: {nn,nnp} 2) lista: longitud 2	1) Palabra: no revelado 2) Sintáctico: jj	1)Semántico: precio

Tabla 3.4. Regla para extraer la cantidad de una transacción sobre una adquisición corporativa.

Por último, en el trabajo realizado por [Téllez A. et al., 2005] se presentó un método para construir un sistema de extracción de información, el cual se denomina *TOPO*. En dicho trabajo de igual manera se efectuaron las fases mostradas en la figura 3.1. Primeramente, durante la fase de *Etiquetado* se identificaron manualmente en un conjunto de 534 noticias sobre desastres naturales, elementos como: fecha, lugar, magnitud, número de muertos, entre otros. Posteriormente, durante la fase de *Pre-procesamiento* se hizo la tarea de Tokenización, donde se identificaron y seleccionaron aquellos fragmentos de textos que eran fuertes candidatos a ser extraídos. Para lograrlo se utilizó un conjunto de expresiones regulares. Después, durante la fase de *Descubrimiento de Patrones de Extracción* se clasificaron los contextos de los fragmentos obtenidos con el fin de identificar su identidad. Es decir, en este caso no se descubrieron patrones para la extracción, sin embargo, el proceso de extracción se planteó como una tarea de clasificación. Por tal motivo, establecieron un clasificador para cada tipo: cantidades, fechas y nombres. Finalmente, durante la fase de *Generación de Plantillas* se utilizaron los clasificadores como medio para extraer la información en los textos dados. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión, recuerdo y medida-F.

De acuerdo con los trabajos descritos anteriormente, se puede concluir lo siguiente:

- Los métodos de extracción realizados bajo el enfoque supervisado requieren de un corpus etiquetado para poder descubrir sus patrones de extracción.
    - La fase de *Etiquetado* requiere de un esfuerzo manual mayor en comparación con el resto de las fases de este tipo de métodos. Debido a que este tipo de métodos requieren de dicha fase, esto ocasiona que sean muy difíciles de llevar a otro dominio. Esto se debe a que para cada dominio es necesario realizar nuevamente la fase de *Etiquetado*.
- Son muy difíciles de llevar a otros lenguajes porque la mayoría de los métodos existentes usan patrones a nivel sintáctico.

Debido a las limitantes con que cuentan este tipo de métodos, surge la idea de los métodos basados en el enfoque no-supervisado.

3.3 Enfoque no-supervisado

Los métodos de extracción de información basados en el enfoque no-supervisado se caracterizan por recibir como entrada un conjunto de documentos no-etiquetados, a partir de dichos textos se descubre un conjunto de patrones de extracción para un dominio específico. Estos patrones sirven para extraer información de textos pertenecientes al mismo dominio.

Este tipo de métodos tienen características similares a los explicados en la sección anterior. Sin embargo, no necesitan de la realización de una fase de etiquetado. Aunque, si es necesaria la intervención de un experto para el etiquetado de los patrones.

En la figura 3.4, se muestra la arquitectura que siguen este tipo de métodos. Esta arquitectura consta de cuatro fases: *Pre-procesamiento, Descubrimiento de Patrones de Extracción, Tipificación de Patrones y Generación de Plantillas.*

La fase de *Pre-procesamiento* que se muestra en la figura 3.4, tiene como objetivo identificar fragmentos de textos de los documentos de entrada. Esta tarea se realiza, ya que es necesario identificar en los textos de entrada los elementos que se pretenden extraer. Los fragmentos de texto resultantes serán útiles para descubrir patrones de extracción a partir de éstos.

Durante la fase de *Pre-procesamiento* se efectúan principalmente algunas de las cuatro tareas siguientes: Tokenización, Análisis Morfológico, Análisis Sintáctico y Reconocimiento de Entidades Nombradas (en la sección 3.2 se explican con más detalle).

Asimismo, la fase de *Descubrimiento de Patrones de Extracción* tiene como objetivo generar un conjunto de patrones de extracción para su posterior tipificación. Para lograr el objetivo, es necesario en algunos casos, dar como entrada algún tipo de información base que ayude al proceso de descubrimiento de patrones. En algunos trabajos se da como entrada: reglas heurísticas, patrones semilla y palabras semilla.

Por otro lado, la fase *Tipificación de Patrones* tiene como objetivo identificar el tipo de información que extraerá cada patrón descubierto con base en una

plantilla dada. Esto es necesario, ya que para la tarea de extracción de información es necesario identificar qué patrón usar para extraer un tipo de información específico. Para poder lograr el objetivo, es necesario realizar un proceso de asignación de tipos a cada patrón descubierto.

Finalmente, la fase *Generación de plantillas* tiene como objetivo generar una plantilla llena por cada documento de entrada, para lograrlo se lleva a cabo la tarea de extracción de información en los textos dados, utilizando los patrones de extracción previamente tipificados en la fase anterior.

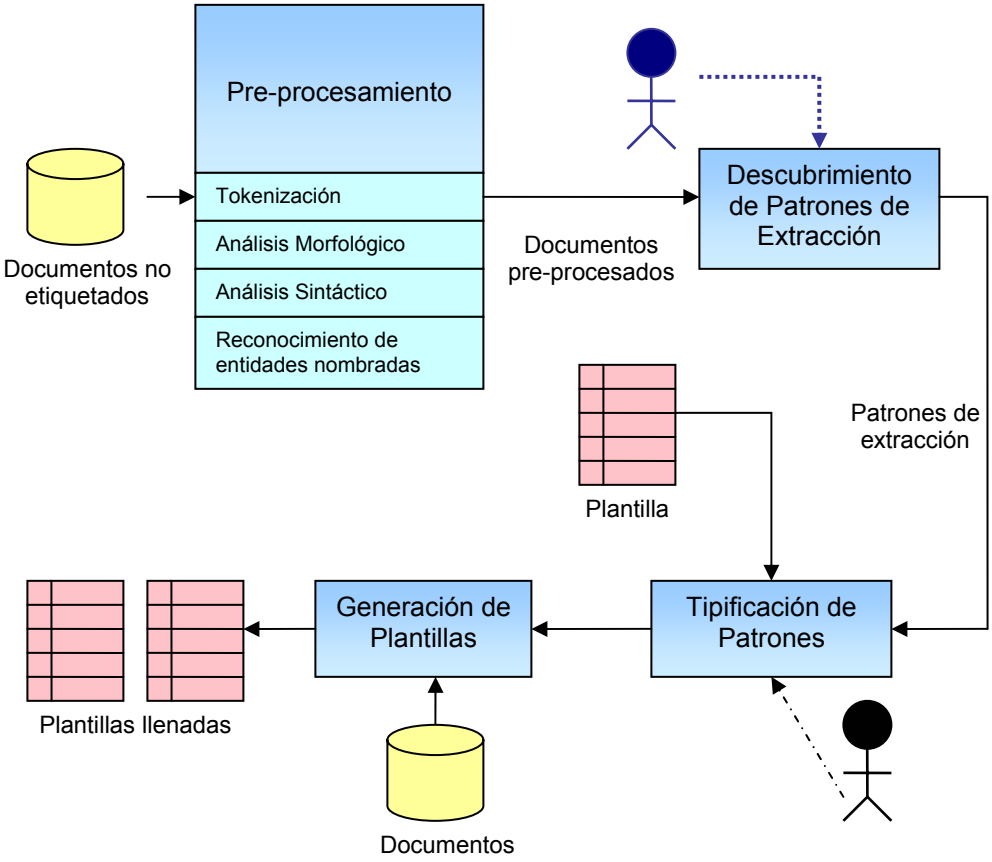

Figura 3.4. Arquitectura de los métodos de extracción de información basados en el enfoque no-supervisado.

Varios de los sistemas de extracción de información desarrollados en los últimos años emplean este enfoque. A continuación, se describirán los trabajos más representativos con respecto a este tema.

En primer lugar, se encuentra el trabajo realizado por [Riloff E., 1996], donde se presentó el sistema AutoSlog-TS. El objetivo de este sistema era crear un diccionario de patrones de extracción útiles para extraer información en textos no etiquetados.

De acuerdo con la arquitectura mostrada en la figura 3.4, en el trabajo de Ellen Riloff se llevaron a cabo las cuatro fases con que cuenta dicha arquitectura. Aunque, cabe resaltar, que los documentos de entrada estaban preclasificados en relevantes e irrelevantes, para lo cual utilizaron los textos de MUC-4, donde cerca del 50% de éstos eran relevantes. Una vez preclasificados los textos, lo siguiente que se realizó fue la fase de *Pre-procesamiento*. Durante esta fase se llevaron a cabo las tareas de Tokenización y Análisis Sintáctico. Posteriormente, durante la fase de *Descubrimiento de Patrones de Extracción* se utilizaron 15 reglas heurísticas establecidas manualmente. Dichas reglas se aplicaron sobre un conjunto de documentos referentes al tema “Atentados Terroristas en América Latina”. El resultado fueron 32,345 patrones sintácticos capaces de extraer cada frase nominal de los textos de entrada. Después, durante la fase de *Tipificación de Patrones* se le calculó a cada patrón un valor de relevancia, de tal forma que los patrones se ordenaron con base en su índice de relevancia. Posteriormente, se llevaron a cabo dos tareas manuales, una de éstas consistió en identificar el tipo de información a extraer de cada patrón, siguiendo el orden de relevancia. La segunda de estas tareas consistió en seleccionar los patrones útiles. Finalmente, durante la fase de *Generación de Plantillas* se emplearon los patrones de extracción resultantes para extraer información sobre “Atentados Terroristas en América Latina”. En la figura 3.5,

se muestra un conjunto de patrones de extracción a nivel sintáctico utilizados en el trabajo de Ellen Riloff. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión, recuerdo y medida-F.

1. <subj> exploded	11. caused <dobj>
2. murder of <np>	12. claimed <dobj>
3. assassination of <np>	13. <subj> was wounded
4. <subj> was killed	14. <subj> occurred
5. <subj> was kidnapped	15. <subj> was located
6. attack on <np>	16. took_place on <np>
7. <subj> was injured	17. responsibility for <np>
8. exploded in <np>	18. occurred on <np>
9. death of <np>	19. was wounded in <np>
10. <subj> took_place	20. destroyed <dobj>

Figura 3.5. Patrones de extracción.

Posteriormente, [Brin S., 1998] presentó el sistema denominado DIPRE. Este sistema empleaba un método bootstrapping[1] para encontrar patrones sin necesidad de dar como entrada documentos etiquetados. De igual manera, en dicho trabajo se llevaron a cabo las cuatro fases explicadas anteriormente. Durante la fase de *Pre-procesamiento* se realizaron las tareas de Tokenización y Reconocimiento de Entidades Nombradas. Después, durante la fase de *Descubrimiento de Patrones de Extracción* fue necesario un conjunto de patrones semilla. Dichos patrones semilla se obtuvieron a partir de las ocurrencias existentes en cinco libros. Dichos libros fueron seleccionados por un experto. El resultado fueron 199 ocurrencias que generaron tres patrones léxicos. El objetivo de dichos patrones era buscar citas de libros. Posteriormente, se aplicaron los patrones obtenidos sobre cinco millones de páginas Web, con el fin de encontrar nuevas ocurrencias

[1] Técnica que consiste en construir primero un modelo con todos los datos iniciales. Entonces, se crean numerosos conjuntos de datos, llamados bootstrap samples, haciendo un muestreo de los datos originales con reemplazo. Posteriormente, se construye un modelo con cada conjunto y se calcula su relación de error sobre el conjunto de prueba.

de los patrones iniciales en textos distintos, mediante dichas ocurrencias se descubrieron nuevos patrones. Posteriormente, se buscaron nuevas ocurrencias en otras páginas Web distintas, mediante el uso de los patrones descubiertos, de tal forma que se volviera un proceso cíclico donde las nuevas páginas Web eran usadas para descubrir más patrones. Enseguida, durante la fase de *Tipificación de Patrones* un experto seleccionó aquellos patrones útiles al proceso. Finalmente, en la fase de *Generación de Plantillas* se utilizaron los patrones descubiertos para extraer pares de autor-título en páginas Web. En la tabla 3.5, se muestran algunos ejemplos de patrones de extracción utilizados en este trabajo. Es importante señalar que cada patrón tiene asignado el url de la página de donde se obtuvo.

URL del patrón	Patrón léxico
www.sff.net/locus/c.*	<LI><B>title</B> by autor (
dns.city-net.com/lmann/awards/hugos/1984.html	title by author (
Dolphin.uppen.edu/dcummins/texts/sf-award.htm	Author || title || (

Tabla 3.5. Patrones de extracción.

Un año después [Riloff E. & Jones R., 1999] presentaron un algoritmo boostrapping multi-nivel. Este algoritmo generaba un léxico semántico y patrones de extracción en forma simultánea, los cuales eran de utilidad para los sistemas de extracción de información. En dicho trabajo se siguieron las cuatro fases explicadas anteriormente. Primeramente, durante la fase de *Pre-procesamiento* se realizó la tarea de Tokenización, para la cual se utilizaron dos colecciones de textos, una de páginas Web empresariales recopiladas por el proyecto WebKB [Craven M. et al.,1998] y otra de noticias de terrorismo. La primera de éstas estaba integrada por 4160 páginas Web, en cambio la segunda colección estaba compuesta por 1500 textos. Después, durante la fase de *Descubrimiento de Patrones de Extracción* fue necesario un conjunto de palabras semilla para la categoría semántica de interés.

Dichas palabras semilla fueron establecidas por un experto en el dominio y las usaron como entrada para la técnica de boostrapping. Esta técnica aprendía los patrones de extracción mediante las palabras semilla de entrada. Posteriormente, esta técnica explotaba los patrones de extracción aprendidos con el fin de identificar más palabras que pertenecieran a la categoría semántica de interés. Dichas palabras servirían para descubrir nuevos patrones, lo cual ocasionaba que este proceso se volviera cíclico. Al final de cada ciclo las mejores extracciones se agregaron a un diccionario semántico temporal. Cabe mencionar que cada patrón estaba ordenado con base a una calificación previamente calculada. Después, durante la fase de *Tipificación de Patrones* se seleccionaron los patrones con mayor calificación. Finalmente, durante la fase de *Generación de Plantillas* se utilizaron los patrones seleccionados para extraer información sobre las categorías semánticas manejadas en este trabajo. En la tabla 3.6, se muestran algunos ejemplos de patrones de extracción. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión y recuerdo.

Patrones de lugar	**Patrones de título**	**Patrones de compañía**	**Patrones de lugar del terrorismo**	**Patrones de armas terroristas**
Offices in <x>	served as <x>	owned by <x>	living in <x>	<x> exploded
Facilities in <x>	became <x>	<x>employed	traveled to <x>	threw <x>
operations in <x>	retired <x>	sold to <x>	parts of <x>	bringing <x>

Tabla 3.6. Patrones de extracción.

De igual forma, en [Yangarber et al., 2000] se presentó un procedimiento para encontrar automáticamente patrones de textos no etiquetados. En dicho trabajo se efectuaron las cuatro fases descritas anteriormente. Durante la fase de *Pre-procesamiento* se realizaron las tareas de Tokenización, Análisis Sintáctico y Reconocimiento de Entidades Nombradas. Después, en la fase

de *Descubrimiento de Patrones de Extracción* se estableció un conjunto de patrones semilla propuestos por un experto, los cuales fueron aplicados sobre 5964 documentos de entrada, es decir, se busco la correspondencia que existía del patrón en cada oración de cada documento, de tal forma que se consideraban como relevantes todos aquellos documentos donde existía la correspondencia de algún patrón. Después, se generalizó cada patrón mediante el reemplazo del ítem léxico por una clase nombre. Enseguida, se seleccionaron aquellos patrones de los documentos relevantes cuya distribución estuviera altamente correlacionada con otros documentos relevantes. El objetivo era encontrar patrones candidatos. En la siguiente fase (*Tipificación de Patrones*) se presentaron los patrones candidatos y sus clases a un experto. La tarea del experto consistía en revisar y seleccionar los patrones relevantes al escenario. Finalmente, los patrones resultantes se utilizaron en la fase de *Generación de Plantillas* para extraer información sobre un corpus de noticias de "Negocios". En la tabla 3.7, se muestra la estructura que deben tener los patrones de extracción en este trabajo, donde C-Compañía y C-persona denotan las clases semánticas que contienen entidades nombradas de los tipos semánticos correspondientes. Por otro lado, C-Asignar denota una clase de verbos en inglés, los cuales son: appoint, elect, promote y name. En cambio, C-renunciar consta de los verbos: resign, depart, quit y step-down. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión y recuerdo.

Sujeto	**Verbo**	**Objeto directo**
C-Compañía C-Persona	C-Asignar C-Renunciar	C-Persona

Tabla 3.7. Estructura de los patrones de extracción.

Posteriormente, en [Sudo K. et al., 2001] se introdujo una representación de patrones basada en árbol. En dicho trabajo se siguieron las cuatro fases presentadas en la figura 3.4. Durante la fase de *Pre-procesamiento* se realizaron las tareas de Tokenización, Análisis Morfológico y Reconocimiento de Entidades Nombradas. Posteriormente, en la fase de *Descubrimiento de Patrones de Extracción* se recuperaron los documentos relevantes para el escenario del conjunto de documentos dado. El resultado fueron 300 documentos relevantes. Después, se seleccionaron 300 sentencias relevantes del conjunto de documentos relevantes, de tal forma que todas las oraciones del conjunto de sentencias relevantes se analizaron mediante un árbol de dependencia. Enseguida, se consideraron todos los predicados del árbol como las raíces para después extraer el camino de la raíz al nodo. Al final, a todos los caminos seleccionados se les calculó un valor de frecuencia. Posteriormente, durante la fase de *Tipificación de Patrones* se seleccionaron aquellos caminos que cumplían con una frecuencia mayor a un umbral dado, y por consiguiente, fueron considerados patrones de extracción, para lograrlo también consideraron el tipo de información del patrón. Por tal motivo, seleccionaron dos escenarios, uno sobre sucesión de directivos y otro sobre detenciones por robo. Un ejemplo de un patrón de extracción utilizado en este trabajo es el siguiente: **<organization>→<post>→appoint**. Finalmente, durante la fase de *Generación de Plantillas* se utilizaron los patrones descubiertos para extraer información sobre los dos escenarios considerados. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión y recuerdo.

Otro de los trabajos que surgieron en este mismo año es el de [Chang C. & Lui S., 2001], donde se presentó un sistema denominado IEPAD. Este sistema descubría automáticamente reglas de extracción de páginas Web.

En el trabajo de Chia-Hui Chang y Shao-Chen Lui de igual manera se realizaron las cuatro fases explicadas al inicio de esta sección. En primer lugar, durante la fase de *Pre-procesamiento* se llevó a cabo la tarea de Tokenización de 140 paginas HTML, donde cada token fue representado como un código binario de longitud fija. Posteriormente, para la fase de *Descubrimiento de Patrones de Extracción* un constructor de árboles PAT recibió cada archivo binario con el fin de construir un árbol PAT (también denominado árbol Patricia [Morrison D., 1968]). Los árboles PAT construidos fueron usados para descubrir patrones repetitivos llamados repeticiones maximales. Dichas repeticiones fueron dadas a un validador. La función de este validador era filtrar los patrones indeseados y producir patrones candidatos. Al final, un compositor de reglas revisaba cada patrón candidato, de tal forma que los patrones candidatos se convirtieran en patrones de extracción con forma de expresiones regulares. Después, en la fase de *Tipificación de Patrones* el usuario del sistema IEPAD seleccionaba manualmente aquel patrón útil para el tipo de información que deseaba extraer. Finalmente, durante la fase de *Generación de Plantillas* el sistema IEPAD mediante un módulo de extracción realizó el proceso de extracción sobre páginas HTML, para lograrlo utilizó los patrones descubiertos en la fase anterior. En la figura 3.6, se muestra un ejemplo de un patrón de extracción usado en este trabajo, así como algunas características propias de dicho patrón. Estas características proporcionan información extra que ayudará al usuario a tomar la decisión de qué patrón utilizar.

1. <DT><STRING></DT><DD><STRING>
<STRING></DD>
Regularidad: 0.000000
Proximidad: 1.000000
Localidad: 1.000000
Riqueza: 0.345456
Longitud: 9
Frecuencia: 15
Identificador: 0

Figura 3.6. Patrón de extracción.

Asimismo, en [Riloff E. et al., 2002] se presentó un método para crear rápidamente sistemas de extracción de información para nuevos lenguajes, para lograr esto, se utilizó la explotación de los sistemas de extracción de información existentes por medio de la proyección-cruzada del lenguaje. En dicho trabajo se realizaron las fases mostradas en la figura 3.4. Durante la fase de *Pre-procesamiento* se llevaron a cabo las tareas de Tokenización y Análisis Sintáctico. Posteriormente, durante la fase de *Descubrimiento de Patrones de Extracción* se utilizó el sistema AutoSlog-TS para generar patrones de extracción para el dominio de accidentes de avión. Dicho sistema generaba una lista de patrones sintácticos ordenados de acuerdo a su asociación con el dominio. En la figura 3.7, se muestran ejemplos de patrones de este tipo. Después, durante la fase de *Tipificación de Patrones* un experto en el dominio revisó los patrones descubiertos y determinó cuales de estos patrones eran útiles para la tarea, para cumplir con esta tarea se consideraron principalmente los patrones situados al principio de la lista ordenada. Finalmente, se utilizaron los patrones seleccionados en la fase de *Generación de Plantillas* para realizar extracción de información sobre textos que pertenecían al dominio de accidentes de avión. Es importante señalar que en este trabajo se presentaron varios experimentos. Estos experimentos mostraban cómo un sistema de extracción de información en inglés podía ser convertido a un sistema en el idioma francés. Sin embargo, esto sólo se realizó para el dominio de accidentes de avión. Cabe mencionar que el corpus usado en dichos experimentos fue obtenido de artículos de periódico escritos en inglés y francés. Este corpus fue creado automáticamente mediante la búsqueda de palabras claves sobre accidentes de avión, de tal forma que el corpus en inglés contenía 420,000 palabras claves, en cambio el de francés contenía sólo 150,000 palabras de ese tipo. Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión, recuerdo y medida-F.

1. <subject >crashed
2. hijacked <direct-object>
3. wreckage of <np>

Figura 3.7. Patrones de extracción para extraer vehículos involucrados en un accidente de avión.

Posteriormente, [Riloff E. et al., 2005] presentaron un sistema de extracción de información que usaba un clasificador de sentencias subjetivas para filtrar las extracciones. El objetivo de este trabajo era que mediante el análisis subjetivo se mejorara la precisión de los sistemas de extracción de información.

De acuerdo con la arquitectura presentada en la figura 3.4, en dicho trabajo se siguieron las cuatro etapas de dicha arquitectura. Primeramente, durante la fase de *Pre-procesamiento* se realizaron las tareas de Tokenización y Análisis Sintáctico de 1400 textos preclasificados en relevantes e irrelevantes. Posteriormente, durante la fase de *Descubrimiento de Patrones de Extracción* el sistema presentado en este trabajo aplicó un clasificador basado en reglas a los textos de entrada. El objetivo de aplicar el clasificador era obtener un conjunto de datos de entrenamiento para generar patrones sintácticos. Un ejemplo de un patrón sintáctico es el siguiente: **<subj>pasive_verb**. En este trabajo utilizaron el sistema AutoSlog-TS. Posteriormente, aplicaron los patrones sintácticos sobre los textos dados, de tal forma que se generaron patrones de extracción para cada instancia de los patrones sintácticos, la idea era aplicar los patrones de extracción a un corpus de entrenamiento y de esa manera obtener un conjunto de estadísticas. Las estadísticas indicaban la relevancia de los patrones ocurridos en los textos relevantes contra la relevancia de éstos en los documentos irrelevantes. Al final, los patrones descubiertos se ordenaron con base en su asociación con el dominio. Enseguida, durante la fase de *Tipificación de Patrones* un experto seleccionó y tipificó manualmente los

patrones útiles para la tarea, para lograr esto, el experto tuvo que analizar 40,553 patrones. Al final se seleccionaron y tipificaron 397 patrones útiles. En la figura 3.8, se muestran algunos ejemplos de patrones de extracción útiles empleados en este trabajo. Finalmente, durante la fase de *Generación de Plantillas,* los patrones descubiertos se utilizaron para extraer información de 200 textos sobre "Atentados Terroristas en América Latina". Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión, recuerdo y medida-F.

1. <subj>was killed
2. <subj> was bombed
3. <subj> was attacked

Figura 3.8. Patrones de extracción útiles.

Por último, se describe el trabajo realizado por [Patward S. & Riloff E.,2006]. En este trabajo se buscaba explorar la idea de usar la Web para identificar de forma automática patrones de extracción de un dominio específico.

De acuerdo con la arquitectura descrita anteriormente, en dicho trabajo se realizaron las cuatro fases con que cuenta dicha arquitectura. Primeramente, durante la fase de *Pre-procesamiento*, se efectuaron las tareas de Tokenización y Análisis Sintáctico sobre 6,182 documentos HTML. Dichos documentos fueron recopilados mediante el uso del buscador Google[2] y 80 preguntas hechas manualmente. Posteriormente, durante la fase de *Descubrimiento de Patrones de Extracción* fue necesario establecer un conjunto de patrones semillas, los cuales fueron establecidos por un experto en el dominio, para identificar oraciones en los textos de entrada, de tal forma que mediante dichas oraciones se pudieran identificar patrones de extracción para el dominio de "Atentados Terroristas en América Latina".

[2] http://www.google.com

Después, aplicaron el sistema AutoSlog-TS a los documentos HTML, con el fin de generar todas las instancias léxicas de los patrones de extracción. Una vez hecho esto, calcularon una correlación estadística de cada patrón con los patrones semilla. Asimismo, aquellos patrones que no ocurrieron en la misma sentencia como un patrón semilla fueron descartados. Enseguida, durante la fase de *Tipificación de Patrones* se cálculo una afinidad semántica de cada patrón candidato con respecto a las categorías: objetivo, victima, criminal, organización, arma y otros. En la figura 3.9, se muestran ejemplos de patrones de extracción usados en este trabajo para identificar los objetivos de un ataque terrorista. Cabe mencionar que para calcular la afinidad semántica de cada patrón sintáctico utilizaron el paquete Sundance [Riloff E. & Phillips W., 2004]. Finalmente, durante la fase de *Generación de Plantillas*, se emplearon los patrones descubiertos para extraer información sobre "Atentados Terroristas en América Latina". Cabe mencionar que el proceso de evaluación en este trabajo se llevó a cabo utilizando las métricas de precisión, recuerdo y medida-F.

1. fired into <np>
2. went off in <np>
3. car bomb near

Figura 3.9. Patrones de extracción.

De acuerdo con los trabajos descritos anteriormente, se puede concluir lo siguiente:

- A pesar del esfuerzo realizado en este tipo de métodos, no se alcanzó a reducir la factibilidad de llevarlos a otros dominios. Esto se debe en gran medida a que los patrones obtenidos son muy complejos y por consiguiente, no cualquier persona puede etiquetarlos.
- Su traslado a otros lenguajes es muy complejo porque en su mayoría utilizan patrones sintácticos.

3.4 Evaluación

Como se puede apreciar en la figura 3.10, el proceso de evaluación de un sistema de extracción de información consiste en comparar claves contra respuestas, es decir, se comparan el conjunto de plantillas manuales generadas por un experto en el dominio (claves) contra el conjunto de plantillas generadas por el sistema de extracción a ser evaluado (respuestas). Para cumplir con esta tarea en [Chinchor N., 1998] utilizan los elementos mostrados en la tabla 3.8, que servirán para calcular las métricas de evaluación definidas en la tabla 3.9.

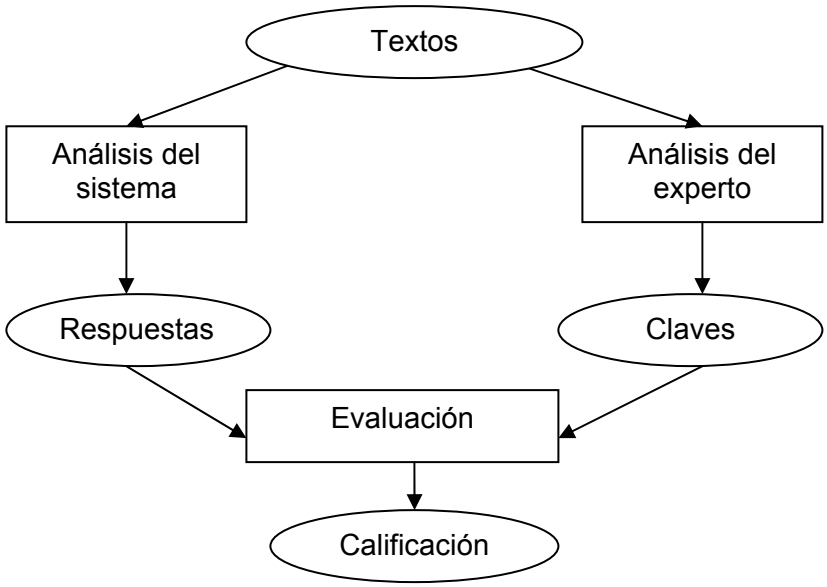

Figura 3.10. Proceso de evaluación de un sistema de extracción de información.

COR	Número correcto	Ocasiones donde la clave y la respuesta coinciden.
INC	Número incorrecto	Ocasiones donde la clave y la respuesta no coinciden.
MIS	Número perdido	Ocasiones donde existe una clave pero no una respuesta.
SPU	Número falso	Ocasiones donde existe una respuesta pero no una clave.
POS=COR+INC+MIS	Número posible	Número de registros en la clave.
ACT=COR+INC+SPU	Número actual	Número de registros en la respuesta.

Tabla 3.8. Elementos usados para calcular las métricas de evaluación.

Recuerdo (R)	$R = \frac{COR}{POS}$
Precisión (P)	$P = \frac{COR}{ACT}$
Medida-F (F)	$F = \frac{2PR}{P+R}$

Tabla 3.9. Métricas usadas para la evaluación de un sistema de extracción de información.

Muchas de las evaluaciones MUC se basan en dar un punto de calificación para cada campo llenado correctamente (COR). En este caso, los campos falsos (SPU) también se cuentan, éstos son campos que son generados y llenados a pesar de que no exista información en el texto. También, están los campos llenados en forma incorrecta (INC) que de igual forma se contabilizan. Por consiguiente, el total de campos correctos en una plantilla también es conocido. Dicha información permite el cálculo de las dos métricas básicas para medir el desempeño de los sistemas de extracción de información y de recuperación de información: *Precisión* (P) y *Recuerdo*(R).

La *Precisión* (P) es definida como la proporción de los campos recuperados y relevantes de entre todos los campos recuperados. En cambio, el *Recuerdo*

(R) se define como la proporción de los campos relevantes que fueron recuperados de entre todos los campos relevantes disponibles.

Posteriormente, debido al interés de tener una medida de evaluación simple, surgió la medida-F (F). Esta medida es definida como la métrica armónica pesada de P y R [Van-Rijsbergen C. ,1979]. Por lo que después, en [Mackhoul J. et al., 1999] esta métrica fue definida como:

$$F = \left[\frac{\alpha}{P} + \frac{1-\alpha}{R} \right]^{-1} = \frac{PR}{(1-\alpha)P + \alpha R} \qquad \text{Donde } 0 \leq \alpha \leq 1. \tag{3.1}$$

Posteriormente, en MUC-6 se estableció que el valor más popular de α era 0.5 y por consiguiente, esta medida se redujo a:

$$F = \frac{2PR}{P+R} \tag{3.2}$$

Por ejemplo, [Riloff E. & Jones R., 1999, Téllez A. et al., 2005, Sabou M., 2005, Riloff E. et al., 2005, Patward S. & Riloff E.,2006] son algunos de los que han utilizado las métricas mostradas en la tabla 3.9, para la evaluación de los sistemas de extracción que proponen.

En este trabajo se utilizaron las métricas de precisión, recuerdo y medida-F para la evaluación del método de extracción de información que se presenta en esta tesis.

Capítulo 4

Método Basado en Patrones de Extracción de Tamaño Variable

En este capítulo se describe la aportación de esta tesis al área de extracción de información. Dicha aportación consiste en un método de extracción de información no-supervisado, cuya característica principal es su portabilidad a otros dominios.

Por lo anterior, durante este capítulo se describe la idea detrás de la arquitectura propuesta (sección 4.1), posteriormente, en la sección 4.2 se detallan cada una de las fases del método de extracción de información propuesto. Asimismo, en la sección 4.3 se presentan los resultados obtenidos al emplear este método sobre un corpus perteneciente al dominio de desastres naturales. Por último, en la sección 4.4, se concluye con una discusión acerca del mismo.

4.1 Introducción

Como se pudo observar en el capítulo 3, existen varios trabajos que se han realizado en el campo de Extracción de Información. Entre ellos, los métodos de extracción de información basados en el enfoque supervisado. Este tipo de métodos requieren de un corpus etiquetado para poder descubrir sus patrones de extracción y además, son difíciles de llevar a otros dominios y lenguajes. Por tal motivo, surgieron los métodos de extracción de información basados en el enfoque no-supervisado, cuyo principal objetivo es tratar de reducir los problemas con que cuentan los métodos anteriores. Sin embargo, este tipo de métodos aún son difíciles de llevar a distintos dominios y lenguajes.

Por lo anterior, en esta tesis se presenta un método de extracción de información no-supervisado que pretende disminuir las limitantes descritas anteriormente. Este método utiliza patrones léxicos de extracción ponderados. Dicho método es más portable en comparación con otros. Esto se logra en gran medida al usar patrones de extracción a nivel léxico, ya que debido a la fácil interpretación de éstos, es mucho más sencillo categorizar los patrones de extracción descubiertos. Además, se usa una técnica de agrupamiento con el fin de reducir el número de instancias que el experto etiquetará, ocasionando que la tarea manual del experto sea fácil y rápida. Además, el experto sólo necesitará tener un conocimiento general sobre el dominio para cumplir con su tarea.

4.2 Método propuesto

El “Método Basado en Patrones de Extracción de Tamaño Variable” que se presenta en esta tesis, está construido siguiendo la arquitectura de los

métodos basados en el enfoque no-supervisado (véase figura 3.4). Esta arquitectura consta de cuatros fases: *Pre-procesamiento, Descubrimiento de Patrones de Extracción, Tipificación de Patrones* y *Generación de Plantillas.*

4.2.1 Fase 1: Pre-procesamiento

El objetivo de esta fase es generar un conjunto de oraciones organizadas por tipo (cantidades, fechas y nombres), a partir de una colección de documentos no etiquetados.

La tarea de esta fase es importante, ya que es indispensable conocer las formas en que se expresa la información a extraer para poder descubrir patrones léxicos.

Para lograr el objetivo de esta fase es necesario realizar las siguientes actividades:

1. Dividir los textos de entrada en oraciones.
2. Realizar la tarea de reconocimiento de entidades nombradas. En esta tarea se identifican en cada una de las oraciones elementos como: cantidades, fechas y nombres.
3. Aplicar un proceso de normalización de las oraciones. Este proceso de normalización consiste en asignar una etiqueta específica en el elemento a extraer, de tal forma que las cantidades se sustituyan por la etiqueta “@CANTIDAD@”, las fechas por “@FECHA@” y los nombres por “@NOMBRE@”.
4. Seleccionar sólo aquellas oraciones que contengan al menos un elemento a extraer (podría ser una cantidad, una fecha o un nombre).

El resultado de esta fase es un conjunto de oraciones organizadas por tipo que incluyan al menos un elemento a extraer. Estas oraciones se utilizarán como información base para la siguiente fase del método propuesto.

Un ejemplo que muestre lo que se lleva a cabo en esta fase se explica a continuación. Primeramente, en la figura 4.1 se muestra una noticia de periódico sobre un incendio.

Se incendia bosque
Por REFORMA
(05 Febrero 2001) Un incendio forestal en la comunidad de Zaragoza, en Oaxaca consume ocho hectáreas de ocotes.

Figura 4.1. Noticia de periódico.

Lo primero que se tiene que realizar es dividir el texto en oraciones. Posteriormente, se deben identificar las cantidades, fechas y nombres que se encuentren en las oraciones. En la noticia que se muestra en la figura 4.1 se identificaron los siguientes elementos.

Cantidades:

- Un
- ocho

Fechas:

- 05 Febrero 2001

Nombres:

- REFORMA
- Zaragoza
- Oaxaca

Posteriormente, se sustituirán aquellas cantidades, fechas y nombres localizadas en el texto por las etiquetas @CANTIDAD@, @FECHA@ y @NOMBRE@ respectivamente, de tal forma que “Un” se sustituirá por @CANTIDAD@ y así sucesivamente con el resto. El resultado serán sólo aquellas oraciones que contengan al menos un elemento a extraer (véase la figura 4.2).

Por @NOMBRE@ @FECHA@ @CANTIDAD@ incendio forestal en la comunidad de @NOMBRE@, en @NOMBRE@ consume @CANTIDAD@ hectáreas de ocotes.

Figura 4.2. Oraciones resultantes.

4.2.2 Fase 2: Descubrimiento de Patrones de Extracción

El objetivo de esta fase es descubrir un conjunto de patrones léxicos, a partir del conjunto de oraciones organizadas por tipo resultantes de la fase anterior.

En este método el descubrimiento de patrones léxicos se basa en la idea de encontrar contextos frecuentes que ocurren alrededor de las entidades nombradas.

Para poder alcanzar el objetivo de esta fase es necesario aplicar una técnica de minería de texto. Dicha técnica ayudará a determinar los patrones de extracción. A continuación se enlistan las actividades a realizar en esta fase:

1. Aplicar una técnica para calcular las Secuencias Frecuentes maximales (SFM) (véase sección 2.1.1) existentes en las oraciones de entrada.
2. Seleccionar automáticamente las SFM que contengan algún elemento a extraer (en este caso serían cantidades, fechas y nombres), es decir, que sean patrones de extracción.
3. Clasificar cada uno de los patrones descubiertos en las categorías de: cantidades, fechas y nombres.

En esta fase ya se cuenta con un conjunto de patrones léxicos útiles para la extracción, sin embargo, aún falta identificar el tipo de información que extraerán con base en una plantilla dada.

Para ejemplificar lo que se lleva a cabo en esta fase, consideremos las SFM que se muestran en la figura 4.3.

```
2 SFM de tamaño 1
        [2] QUEDARON
        [2] LAS
1 SFM de tamaño 2
        [2] POR EL
3 SFM de tamaño 3
        [2] @CANTIDAD@ DE MUERTOS
        [2] EN LA REGION
        [2] MURIERON @CANTIDAD@ PERSONAS
2 SFM de tamaño 4
        [2] @CANTIDAD@ HOGARES FUERON AFECTADOS
        [2] SE DERRUMBARON @CANTIDAD@ CASAS
1 SFM de tamaño 5
        [2] EN @NOMBRE@ HUBO UN INCENDIO
```

Figura 4.3. Conjunto de SFM.

Lo que se busca en esta fase es identificar en aquellas SFM los patrones léxicos de extracción, por tal motivo, es necesario identificar las SFM resultantes que contengan algún elemento a extraer, por lo tanto, los patrones léxicos de extracción resultantes a partir de las SFM de la figura 4.3

son los que se muestran en la figura 4.4. Como se puede observar, los patrones están organizados por tipo, cabe notar que en este ejemplo no se obtuvieron patrones para fechas.

CANTIDADES
Patrón 1: @CANTIDAD@ DE MUERTOS
Patrón 2: MURIERON @CANTIDAD@ PERSONAS
Patrón 3: @CANTIDAD@ HOGARES FUERON AFECTADOS
Patrón 4: SE DERRUMBARON @CANTIDAD@ CASAS

NOMBRES
Patrón 5: EN @NOMBRE@ HUBO UN INCENDIO

Figura 4.4. Patrones léxicos resultantes.

4.2.3 Fase 3: Tipificación de Patrones

El objetivo de esta fase es generar un conjunto de patrones léxicos ponderados específicos para un dominio, para ello, es necesario contar con la plantilla de datos a extraer.

Aunque, los patrones léxicos descubiertos en la fase anterior están preclasificados en patrones que extraen cantidades, fechas y nombres, aún no es posible utilizarlos. Esto se debe a que no se sabe qué tipo de cantidad, fecha y nombre extraerán cada uno. Por ejemplo, para el tema de desastres naturales se podrían tener patrones de cantidad para extraer la cantidad de muertos, heridos y desaparecidos. Por tal motivo, aún no es posible emplear los patrones para extraer información, ya que para poder utilizar dichos patrones, es necesaria la intervención de un usuario con conocimientos sobre el área de extracción. La función de éste será asignar el tipo de información que extraerá un conjunto de patrones, con base en la plantilla de datos a extraer. Finalmente, el resultado de esta fase es un conjunto de patrones léxicos de extracción ponderados.

Los patrones léxicos de extracción ponderados tienen asignado un porcentaje por tipo de información. En esta fase se manejan este tipo de patrones, porque los patrones descubiertos en la fase anterior son generales, ocasionando que un mismo patrón pudiera extraer información útil de dos o más campos de la plantilla. Este hecho provocó que fuera necesario realizar un proceso de ponderación de los patrones.

En la figura 4.5, se muestra la arquitectura que se siguió en esta fase para lograr el objetivo. Dicha arquitectura consta de cuatro actividades: *Extracción de Instancias, Agrupamiento de Instancias, Tipificación de Centroides* y *Cálculo de Pesos.*

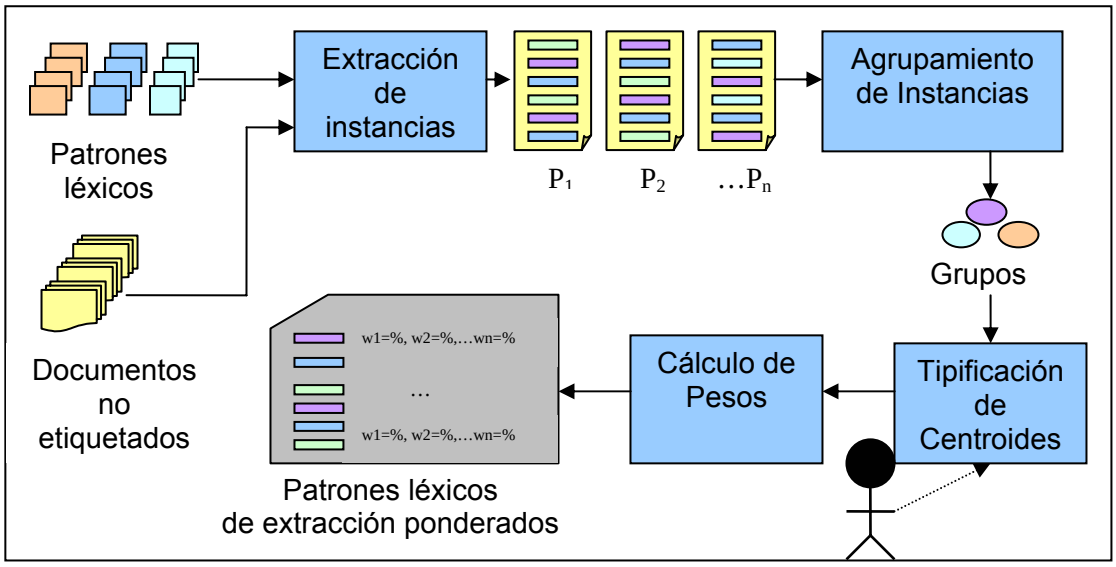

Figura 4.5. Arquitectura de la fase de Tipificación de Patrones.

4.2.3.1 Extracción de Instancias

El objetivo de esta actividad es generar un conjunto de instancias, a partir de la aplicación de los patrones léxicos descubiertos en la fase dos, sobre una colección de documentos no etiquetados.

Para llevar a cabo esta actividad se debe realizar lo siguiente:

1. Buscar en los textos de entrada todas aquellas oraciones (instancias) en las que exista una correspondencia del patrón que se esté aplicando en la oración, de tal forma que el resultado será un conjunto de instancias por patrón.
2. Identificar en cada una de las instancias las cantidades, fechas y nombres existentes.
3. Filtrar las instancias resultantes, de tal forma que el resultado final sean sólo instancias que contengan sólo un elemento a extraer del mismo tipo. Esto se lleva a cabo para evitar la confusión al asignar el tipo de información que extraerá cada patrón.

Un ejemplo que explique el funcionamiento de esta actividad se muestra a continuación.

Primeramente, se debieron haber descubierto un conjunto de patrones léxicos. Por ejemplo, consideremos los patrones léxicos para cantidades que se muestran en la figura 4.4. Posteriormente, se buscan aquellas oraciones en los textos no etiquetados donde exista una correspondencia de cada patrón. Considérese el texto no etiquetado mostrado en la figura 4.6 para buscar la correspondencia de los patrones. Las frases seleccionadas se denominarán instancias por patrón (véase la figura 4.7).

Incendios forestales
Por REFORMA
(31 Marzo 2002).-
En Hidalgo hubo un incendio que ayer consumió algunas comunidades y hasta el momento continúa fuera de control. Se dice que murieron 15 personas en los alrededores. Veinte hogares fueron afectados en el centro. Sin embargo, en las afueras de la ciudad se derrumbaron 10 casas. Según informes murieron 35 personas en total.

Figura 4.6. Noticia de periódico.

Patrón2_1: Se dice que murieron 15 personas en los alrededores.
Patrón2_2: Según informes murieron 35 personas en total.
Patrón3_1: Veinte hogares fueron afectados en el centro.
Patrón4_1: Sin embargo, en las afueras de la ciudad se derrumbaron 10 casas.

Figura 4.7. Instancias resultantes de aplicar los patrones léxicos.

Dichas instancias se identifican por el número del patrón seguido por el número de la instancia (por ejemplo PatrónX_Y significa que es la instancia Y del patrón X). Posteriormente, se identificarán en cada una de las instancias de la figura 4.6, las cantidades, fechas y nombres (véase la figura 4.8), para después, seleccionar sólo aquellas instancias que contengan un elemento a extraer.

Patrón2_1: Se dice que murieron @CANTIDAD@ personas en los alrededores.
Patrón2_2: Según informes murieron @CANTIDAD@ personas en total.
Patrón3_1: @CANTIDAD@ hogares fueron afectados en el centro.
Patrón4_1: Sin embargo, en las afueras de la ciudad se derrumbaron @CANTIDAD@ casas.

Figura 4.8. Instancias etiquetadas.

En este caso las instancias resultantes de esta actividad son las mismas cuatro que se muestran en la figura 4.8. Esto se debió a que ninguna de las instancias generadas contenía dos elementos a extraer del mismo tipo y por tal motivo, ninguna de éstas se filtró.

4.2.3.2 Agrupamiento de Instancias

El objetivo de esta actividad es generar un conjunto de grupos a partir de las instancias. La idea es que cada grupo estará asociado a un tipo específico de información.

En este punto se decidirá si un patrón es útil y el tipo de instancias que extrae. Para ello, se utilizarán las instancias obtenidas anteriormente.

Esta actividad ayudará a reducir el trabajo del experto, es por ello que es necesario agrupar las instancias provenientes de los patrones. De esto dependerá la cantidad de ejemplos que el experto tendrá que etiquetar, ya que éste sólo etiquetará el elemento más representativo de cada grupo.

Para realizar la tarea de agrupamiento, es necesario contar con un algoritmo de este tipo. Este algoritmo deberá tener las siguientes características: (1) que no necesite como entrada el número de grupos a formar y (2) que proporcione la instancia más representativa de cada grupo (centroide), es decir, una instancia por grupo. Es por esto que se empleó el algoritmo estrella (el funcionamiento y características de este algoritmo se describen más a fondo en la sección 2.1.2.1).

Para lograr el objetivo de esta actividad se deben agrupar las instancias resultantes de la actividad anterior (E*xtracción de instancias*) mediante el algoritmo de agrupamiento estrella. El resultado es un conjunto de grupos de instancias con características similares entre ellas.

Para ejemplificar esta fase consideremos las instancias que se muestran en la figura 4.8. Después de agrupar estas instancias mediante el algoritmo estrella tomando un umbral, el resultado sería el que se muestra en la tabla 4.1.

Grupo	Centroide	Satélite
1	**Patrón2_1**	**Patrón2_2**
2	**Patrón3_1**	
3	**Patrón4_1**	

Tabla 4.1. Grupos resultantes.

4.2.3.3 Tipificación de Centroides

El objetivo de esta actividad es asignar el tipo de información a la que se refieren las instancias obtenidas aplicando los patrones léxicos. Esto se logra a partir del etiquetamiento manual de los centroides de cada grupo. Mediante este proceso de etiquetamiento se asumirá que los otros elementos del grupo pertenecen al mismo tipo.

Los pasos a seguir en esta actividad son los siguientes:

1. Identificar las instancias más representativas (centroides) de cada grupo creado durante la actividad anterior (*Agrupamiento de instancias*), en este caso una de cada grupo. Es importante señalar que el algoritmo utilizado (estrella) en esta tesis proporciona los centroides como parte de su funcionamiento.
2. Eliminar aquellos grupos integrados por una sola instancia. Esto es necesario ya que de lo contrario se contaría con grupos muy particulares que no ayudarían al proceso.
3. Determinar en forma manual el tipo de información a la que se refiere cada centroide, respecto a una plantilla específica.
4. Dar a todos los elementos del grupo el tipo asignado al centroide. El objeto de esto es reducir la cantidad de ejemplos a etiquetar en forma manual, lo cual ayudará a establecer el tipo de información que extraerá cada patrón léxico, logrando que el proceso de tipificación de patrones sea más fácil y rápido.

Por ejemplo, los grupos que se muestran en la tabla 4.1 tienen cada uno de éstos un centroide. Cada uno de los centroides se etiquetará manualmente, de acuerdo con el tipo de información a la que se refiere la instancia. En la tabla 4.2, se muestran los tipos asignados únicamente a los centroides.

Aunque la instancia **Patrón2_2** no se etiquetó manualmente, ésta adquiere el tipo asignado al centroide de su grupo. Por lo tanto, la instancia **Patrón2_2** será del tipo "Personas muertas".

Grupo	Centroide	Tipo asignado al centroide	Satélite
1	**Patrón2_1**	Personas muertas	**Patrón2_2**
2	**Patrón3_1**	Viviendas afectadas	
3	**Patrón4_1**	Viviendas destruidas	

Tabla 4.2. Centroides etiquetados.

4.2.3.4 Cálculo de Pesos

El objetivo de esta actividad es generar un conjunto de patrones léxicos de extracción ponderados a partir de una serie de instancias de dichos patrones previamente tipificadas.

Esta actividad se llevó a cabo, porque un patrón podría haber recuperado instancias de distintos tipos, es por ello, que fue necesario asignar una ponderación con respecto al tipo de información que pudiera extraer cada patrón.

Para lograr el objetivo de este paso, es necesario realizar las siguientes tareas:

1. Identificar para cada patrón el número de instancias tipificadas propias del patrón en turno, así como el tipo asignado a cada una de estas instancias.

2. Asignar un porcentaje por tipo a cada patrón. Este porcentaje indica la cobertura que tiene cada patrón con respecto a la cantidad de instancias resultantes. Dicho porcentaje se obtiene a partir de la cantidad de instancias por categoría entre el total de instancias tipificadas por patrón.

El resultado de esta actividad es un conjunto de patrones léxicos de extracción ordenados por su porcentaje y clasificados por categoría. Es por ello que en este caso un patrón podrá extraer información de distintos tipos.

La idea de tener patrones ordenados por un porcentaje, es tener una base que indique cuáles serían los patrones que por su cobertura tienen más posibilidad de extraer la información requerida.

Por ejemplo, si consideramos las instancias etiquetadas de la tabla 4.2, el resultado sería el que se muestra en la tabla 4.3.

Patrón Léxico	Tipo	Porcentaje
Patrón2: MURIERON @CANTIDAD@ PERSONAS	Personas muertas	100%
Patrón 3: @CANTIDAD@ HOGARES FUERON AFECTADOS	Viviendas afectadas	100%
Patrón 4: SE DERRUMBARON @CANTIDAD@ CASAS	Viviendas destruidas	100%

Tabla 4.3. Patrones léxicos de extracción ponderados.

En este caso cada patrón extraerá información de sólo un tipo, sin embargo, existen ocasiones donde el patrón pudiera extraer información de distintos tipos.

Otro ejemplo se muestra en tabla 4.4. En esta tabla se presenta un conjunto de patrones léxicos para extraer información en textos sobre desastres

naturales, así como sus correspondientes porcentajes. En este caso el único patrón que extraerá información de sólo un tipo es el número 1. Este patrón sólo podrá extraer información sobre el número de muertos en un desastre natural. Sin embargo, el resto de los patrones se podrán usar para extraer información de tipos distintos, pero con la posibilidad de que no pueda extraer la información correcta, debido a que cubre un porcentaje menor de las instancias resultantes.

Patrón Léxico	**No. de muertos**	**No. de heridos**	**No. de desaparecidos**	**No. de viviendas destruidas**	**No. de viviendas afectadas**
Patrón 1: LOS @CANTIDAD@ MUERTOS	100%	0%	0%	0%	0%
Patrón 2: @CANTIDAD@ DECESOS	95%	5%	0%	0%	0%
Patrón 3: DEJÓ @CANTIDAD@ PERSONAS	0%	10%	90%	0%	0%
Patrón 4: RESULTARON @CANTIDAD@ VIVIENDAS	0%	0%	0%	66%	34%
Patrón 5: MÁS DE @CANTIDAD@ PERSONAS	25%	75%	0%	0%	0%
Patrón 6: @CANTIDAD@,VOLUNTARIOS	25%	50%	25%	0%	0%

Tabla 4.4 Patrones léxicos de extracción ponderados.

De acuerdo con la tabla 4.4, los mejores patrones para extraer el número de muertos son los patrones 1 y 2. Por otro lado, los patrones que mejores resultados arrojarán si se utilizan para extraer el número de heridos son los patrones 5 y 6. De igual manera, el mejor patrón para extraer el número de desaparecidos en un desastre natural es el patrón 3. Finalmente, para extraer el número de viviendas afectadas y destruidas sólo se podrá emplear el patrón 4. Este patrón fue el único que se tipificó como de estos tipos.

4.2.4 Fase 4: Generación de Plantillas

El objetivo de esta fase es generar un registro o plantilla por cada documento de entrada. Cabe mencionar que la información a extraer será establecida desde el inicio del proceso. Dicha información dependerá de los datos que sean de interés sobre el dominio a utilizar.

Para lograr el objetivo de esta fase es necesario aplicar los patrones léxicos de extracción ponderados descubiertos en la fase anterior en el orden en que se encuentran. Dichos patrones se aplicarán exclusivamente sobre textos que pertenezcan al dominio de extracción.

Por ejemplo, en la tabla 4.4 se muestra una serie de patrones léxicos ponderados para extraer cantidades en desastres naturales. El proceso de extracción de información se debe realizar aplicando los patrones léxicos con base en su porcentaje. Es decir, se deberá aplicar primeramente el patrón con mayor porcentaje de acuerdo al tipo de información a extraer. Esto se debe a que el patrón con mayor porcentaje tiene una alta posibilidad de extraer correctamente la información. Posteriormente, se deberá aplicar el resto de los patrones en el orden establecido, sólo si no se extrajo la información correcta. El resultado final en esta fase será una plantilla por cada documento de entrada. Dicha plantilla contendrá la información que se desea obtener de cada texto dado.

4.3 Resultados Experimentales

Para la evaluación del método de extracción propuesto se empleó el corpus de desastres naturales utilizado en [Téllez A., 2005]. El tipo de información que se maneja en este dominio se muestra en la tabla 4.5. Asimismo, en la

investigación de [Téllez A., 2005] se propone un método de extracción de información basado en el enfoque supervisado denominado TOPO.

CANTIDADES	
PER_MUERTAS	Número de personas fallecidas por causas directas.
PER_HERIDAS	Número de personas que resultan afectadas en salud o integridad física, sin ser victimas mortales, por causa directa del desastre.
PER_DESAPARECIDAS	Número de personas cuyo paradero a partir del desastre es desconocido.
PER_DAMNIFICADAS	Número de personas que han sufrido grave daño directamente asociados al evento en sus bienes o servicios.
PER_AFECTADAS	Número de personas que sufren de efectos secundarios asociados a un desastre.
VIV_DESTRUIDAS	Número de viviendas arrasadas, sepultadas, colapsadas o deterioradas de tal manera que no son habitables.
VIV_AFECTADAS	Número de viviendas con daños menores, no estructurales o arquitectónicos, que pueden seguir siendo habitadas.
INF_HECTAREAS	Número de áreas de cultivo, pastizales o bosques destruidos o afectados.
INF_ECONOMICA	Monto de las pérdidas directas causadas por el desastre.
FECHAS	
EVE_FECHA	Fecha de ocurrencia del desastre.
NOMBRES	
EVE_LUGAR	Nombre del lugar o lugares donde ocurrió el fenómeno.

Tabla 4.5. Tipos de información manejada en el dominio de desastres.

Los experimentos realizados en esta tesis se compararon con los resultados obtenidos por TOPO.

Para realizar los experimentos se utilizaron varios conjuntos de noticias de periódicos, sin embargo, se presentan los resultados de los experimentos que mejores resultados dieron con un conjunto específico de noticias, el cual estaba integrado por 550 noticias.

De acuerdo con la arquitectura del método de extracción de información que se detalla en este trabajo (véase figura 3.4), ésta consta de cuatro fases, de tal forma que para la fase 1, se utilizaron 500 noticias de periódicos sobre desastres naturales. De igual manera, para la fase 3 se utilizaron las mismas 500 noticias de la fase 1. Por último, en la fase 4 se realizó el proceso de evaluación utilizando 50 noticias de periódicos del mismo dominio.

El primer experimento que se llevó a cabo en esta tesis tenía como objetivo probar el método de extracción no-supervisado propuesto en un dominio sobre desastres naturales.

A continuación, se detallan los resultados obtenidos durante la realización de este experimento.

Durante la fase "*Descubrimiento de Patrones de Extracción*" se obtuvieron un conjunto de patrones léxicos descubiertos a partir de las SFM calculadas. Para calcular dichas SFM se determinó de manera empírica el umbral con valor de 5. En la tabla 4.6, se muestra la cantidad de patrones léxicos encontrados en esta fase.

Cantidad de patrones léxicos	**Cantidad de patrones léxicos resultantes después del filtrado**
CANTIDADES	
465	459
FECHAS	
11	8
NOMBRES	
414	367

Tabla 4.6. Total de patrones léxicos obtenidos en la Fase 2.

Como se puede ver en la tabla 4.6, la cantidad de patrones léxicos que se muestran en la primera columna se obtuvieron a partir de las SFM

calculadas. Sin embargo, a estos patrones se les tuvo que aplicar un proceso de filtrado para evitar confusión en la tarea de extraer la información, especialmente, en los casos donde los patrones estaban conformados por dos o más elementos a extraer del mismo tipo. En la segunda columna de la tabla 4.6, se muestra la cantidad de patrones léxicos resultantes después de realizar el filtrado.

De acuerdo con las cantidades mostradas en la tabla 4.6, se puede concluir lo siguiente:

- La pérdida de patrones léxicos durante la fase 2 fue mínima. Esto se debe a que se mantuvo más del 93% de los patrones a pesar del proceso de filtrado.

Durante la fase "*Tipificación de Patrones*", se utilizaron los patrones léxicos descubiertos durante la fase "*Descubrimiento de Patrones de Extracción*". Como se explicó en la sección 4.2.3.1, dichos patrones léxicos se aplicaron sobre 500 noticias de desastres naturales. Posteriormente, las instancias resultantes de aplicar los patrones léxicos se filtraron, de tal forma que sólo se dejaron aquellas instancias con un elemento a extraer del mismo tipo. Los resultados se muestran en la tabla 4.7.

No. patrones aplicados	**No. inicial de instancias**	**No. final de instancias**	**Promedio: no. instancias x patrón**
CANTIDADES			
459	3842	547	1.19
FECHAS			
8	11	1	0.12
NOMBRES			
367	1505	193	0.52

Tabla 4.7. Número total de instancias generadas.

De la tabla 4.7, se destaca lo siguiente:

- Para la clase cantidades se generaron 3842 instancias, de las cuales sólo se mantuvieron 547 después del filtrado. Es decir, se eliminó el 85.76% de las instancias generadas.
- Para la clase fechas se generaron 11 instancias, de las cuales sólo se mantuvo una después del filtrado. Es decir, se eliminó el 90.9% de las instancias generadas.
- Finalmente, para la clase nombres se generaron 1505 instancias, de las cuales se mantuvieron 193 después del filtrado. Es decir, se eliminó el 87.17% de las instancias generadas.

Por lo tanto, la cantidad de instancias resultantes para utilizarse en la actividad "*Agrupamiento de Instancias*" de la fase 3, es menor en comparación con las generadas inicialmente. Como se verá, en un experimento posterior, se buscó un medio para abordar este problema.

A pesar de contar con una pequeña cantidad de instancias, éstas se utilizaron en la actividad siguiente (*Agrupamiento de Instancias)*. Durante esta actividad se agruparon las instancias finales mediante el algoritmo de agrupamiento estrella. Como se vió en la sección 2.1.2.1, es necesario identificar un umbral, para lo cual se consideraron dos umbrales, uno que permitiera ser más laxo para formar los grupos y otro más estricto para formar los grupos. Es por esto que se eligieron los umbrales μ y $\mu+\sigma$ respectivamente, de tal forma que al usar el μ se obtuvieron menos grupos pero con una mayor cantidad de elementos. En cambio, al usar $\mu+\sigma$ se obtuvieron más grupos pero con una menor cantidad de elementos en comparación con el umbral anterior. Cabe mencionar, que el umbral μ es el valor promedio de las similitudes de las instancias. En cambio, el otro umbral

además del promedio utiliza la desviación estándar. Es importante mencionar que el umbral que dió mejores resultados fue el de $\mu + \sigma$. Por tal motivo, sólo se presentan los resultados que se obtuvieron utilizando dicho umbral. Para recordar el cálculo de μ y $\mu + \sigma$ se dan las siguientes ecuaciones.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ tal que } x_i \in X \tag{4.1}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(\mu - x_i)}{n}} \text{ tal que } x_i \in X \tag{4.2}$$

Asimismo, como se explicó en la sección 4.2.3.3, la siguiente actividad fue etiquetar los centroides de los grupos formados. En la tabla 4.8, se muestra la cantidad de centroides etiquetados en este experimento.

Clases	No. total de centroides etiquetados
CANTIDADES	169
FECHAS	1
NOMBRES	102

Tabla 4.8. Centroides etiquetados manualmente.

Categorías	No. total de patrones ponderados
CANTIDADES	
EVE_MAGNITUD	7
PER_MUERTAS	95
PER_HERIDAS	1
PER_DESAPARECIDAS	1
PER_DAMNIFICADAS	4
PER_AFECTADAS	32
VIV_DESTRUIDAS	8
VIV_AFECTADAS	9
INF_HECTAREAS	11
INF_ECONOMICA	3
FECHAS	
EVE_FECHA	7
NOMBRES	
EVE_LUGAR	46

Tabla 4.9. Número de patrones léxicos de extracción resultantes.

Con base en lo anterior, en la tabla 4.9 se muestran la cantidad de patrones léxicos de extracción obtenidos como resultado de la fase 3. De esta tabla se destaca que la cantidad de patrones para extraer información del tipo PER_HERIDAS, PER_DESAPARECIDAS e INF_ECONOMICA no es suficiente para abarcar la generalidad de la información sobre este tipo.

Finalmente, durante la fase "*Generación de Plantillas*" se realizó el proceso de extracción de información, cuya plantilla de salida fue la que se muestra en la figura 4.9. Para lograr la realización de esta fase se aplicaron los patrones léxicos de extracción resultantes a 50 noticias de desastres naturales. Estas noticias fueron las mismas con que se evaluó el TOPO. Al final, se evaluó el método de extracción propuesto con base en las medidas de precisión, recuerdo y medida-F (véase tabla 4.10).

INFORMACIÓN DEL DESASTRE	
Fecha	
Lugar	
Magnitud	
INFORMACIÓN DE LAS PERSONAS	
Muertos	
Heridos	
Desaparecidos	
Damnificados	
Afectados	
INFORMACIÓN DE LAS VIVIENDAS	
Destruidas	
Afectadas	
INFORMACIÓN DE LA INFRAESTRUCTURA	
Hectáreas	
Económica	

Figura 4.9. Plantilla de extracción para el dominio de Desastres Naturales.

Evaluaciones	**Precisión**	**Recuerdo**	**Medida-F**
TOPO	0.809	0.885	0.845
Método propuesto	0.715	0.751	0.733

Tabla 4.10. Resultados de la Evaluación.

De acuerdo con los resultados de este experimento se puede destacar lo siguiente:

- Los patrones léxicos obtenidos en general son pocos, considerando que se está trabajando a un nivel léxico, esto ocasiona baja cobertura.
- El sistema TOPO superó al método de extracción propuesto en este trabajo. Sin embargo, el método propuesto es más rápido de construirse a diferencia del TOPO. Además, en el método propuesto se etiqueta una mínima muestra del conjunto total de ejemplos a etiquetar. Por ejemplo, para este experimento se tuvieron que etiquetar únicamente 272 oraciones, en cambio el TOPO tendría que haber etiquetado aproximadamente 6600 oraciones.
- Este método realiza la tarea de extracción de información mediante el descubrimiento automático de patrones a nivel léxico, haciendo más fácil el proceso de llevarlo a otro dominio.

Analizando los resultados se observó que en este experimento se filtraron muchas instancias. Estas instancias se agrupan mediante el algoritmo estrella. El hecho de que existan pocas instancias provoca que haya menos grupos y por consiguiente, se reduzca la cantidad de patrones. Es por esto que fue necesario realizar un segundo experimento, en donde no se filtren las instancias, con el fin de incrementar la cantidad de patrones.

El objetivo del segundo experimento de esta tesis es probar el método de extracción no-supervisado propuesto en un dominio sobre desastres naturales, sin filtrar las instancias.

Para este experimento, se emplearon las instancias obtenidas inicialmente del experimento anterior, lo cual significa que para este experimento no se

filtraron las instancias. En la tabla 4.11, se muestra el número de instancias que se utilizaron en este experimento.

No. patrones aplicados	No. inicial de instancias	No. final de instancias	Promedio: no. instancias x patrón
CANTIDADES			
459	3842	3842	8.37
FECHAS			
8	11	11	1.37
NOMBRES			
367	1505	1505	4.10

Tabla 4.11. Número total de instancias.

Después, durante la fase "*Tipificación de Patrones*" se utilizó una mayor cantidad de instancias en comparación con el experimento anterior. Asimismo, durante la actividad de "*Agrupamiento de Instancias*" se aumentó la cantidad de centroides a etiquetar (véase tabla 4.12) en comparación con el primer experimento.

Clases	No. Total de centroides etiquetados
CANTIDADES	177
FECHAS	8
NOMBRES	147

Tabla 4.12. Centroides etiquetados manualmente.

Al aumentar la cantidad de instancias durante la fase 3, también se incrementó la cantidad de patrones léxicos de extracción ponderados resultantes de esta fase (véase la tabla 4.13).

Categorías	No. Total de patrones ponderados
CANTIDADES	
EVE_MAGNITUD	11
PER_MUERTAS	83
PER_HERIDAS	6
PER_DESAPARECIDAS	7
PER_DAMNIFICADAS	28
PER_AFECTADAS	164
VIV_DESTRUIDAS	15
VIV_AFECTADAS	35
INF_HECTAREAS	45
INF_ECONOMICA	14
FECHAS	
EVE_FECHA	2
NOMBRES	
EVE_LUGAR	189

Tabla 4.13. Número de patrones léxicos de extracción resultantes.

De la tabla 4.13 se puede destacar lo siguiente:

- El número de patrones léxicos aumentó de 224 a 599.

Posteriormente, se realizó el proceso de extracción de información, ahora con un mayor número de patrones léxicos. En la tabla 4.14, se muestran los resultados de la evaluación generados para este experimento (No.2) y para el experimento anterior (No.1).

Evaluaciones	Precisión	Recuerdo	Medida-F
TOPO	0.809	0.885	0.845
No.1	0.715	0.751	0.733
No.2	0.646	0.767	0.701

Tabla 4.14. Resultados de la evaluación.

En los resultados de evaluación se pudo observar lo siguiente:

- Se aumentó el recuerdo de un 75.1% a un 76.7%. Esto se debe a que al haber un mayor número de patrones léxicos se extrajeron más elementos, entre los cuales había una mayor cantidad de resultados correctos en comparación con el experimento anterior.
- Se redujo la precisión de un 71.5% a un 64.6%. Esto se debe a que al haber una mayor cantidad de patrones léxicos se extrajeron más elementos que en su mayoría eran datos incorrectos, en comparación con el experimento anterior.

Dados los resultados anteriores, se determinaron los máximos de precisión y recuerdo que se podían alcanzar con la información que se tenía. Por tal motivo, se agruparon manualmente los patrones resultantes del experimento 2 (véase tabla 4.13) con el objeto de determinar qué tarea tenía mayor impacto (el agrupamiento o la tipificación). Posteriormente, se emplearon los patrones agrupados manualmente en el proceso de extracción. Los resultados se muestran en la tabla 4.15.

Evaluaciones	**Precisión**	**Recuerdo**	**Medida-F**
TOPO	0.809	0.885	0.845
Manual	0.746	0.777	0.761

Tabla 4.15. Resultados de la evaluación.

De la tabla 4.15 se puede destacar lo siguiente:

- El método de extracción de información propuesto sólo puede alcanzar un 76.1% de medida-F como máximo, empleando un corpus de 500 noticias. Por lo tanto, el principal problema con este método no radica ni en el agrupamiento ni en la tipificación, sino en la cantidad de

información de entrada utilizada para generar patrones léxicos de extracción.

4.4 Discusión

Con base en los resultados mostrados en la sección anterior, se pudo determinar lo siguiente:

- El método de extracción de información propuesto puede alcanzar un 76.1% de medida-F como máximo.
- Se tienen pocos patrones léxicos para el proceso de extracción. Cómo se vió en la tabla 4.14, este hecho ocasionó una baja cobertura.
- Como se vió en la evaluación manual (véase tabla 4.15), este método requiere de una mayor cantidad de información para poder generar más patrones léxicos.
- La técnica de minería de texto utilizada es muy exigente. Por tal motivo, no es adecuado usarla si lo que se busca es tener la mayor cantidad de información posible.

Es por esto que es indispensable descubrir los patrones léxicos de una forma distinta a la que se realizó. Es decir, que no se utilice una técnica de minería de texto para descubrir los patrones léxicos. Esta nueva forma de realizarlo deberá permitir contar con la mayor cantidad de información posible para poder generar los patrones léxicos, logrando con ello aumentar la cantidad de patrones léxicos de extracción, de tal forma que al tener más patrones léxicos sea posible incrementar los porcentajes de precisión y recuerdo.

En el próximo capítulo, se presenta un nuevo método tratando de resolver el problema descrito anteriormente.

Capítulo 5

Método Basado en Patrones de Extracción de Tamaño Fijo

En este capítulo se describe el “Método Basado en Patrones de Extracción de Tamaño Fijo”. Este nuevo método se propone dado que se infiere que no se está descubriendo la cantidad suficiente de patrones léxicos. Se considera que usar Secuencias Frecuentes Maximales (SFM) es muy restrictivo, ya que filtran los textos de entrada y por consiguiente, reducen la cantidad de información utilizada en el método.

Por lo anterior, se plantea este nuevo método, el cual se caracteriza por utilizar los contextos directamente de los textos de entrada, de tal forma que se tenga una mayor cantidad de información en comparación con el método anterior. Es por esto que ya no se calcularán las SFM a partir de las oraciones de entrada.

En este capítulo se describe primeramente la idea detrás de la nueva arquitectura propuesta (véase la sección 5.1). Posteriormente, en la sección 5.2 se detallan cada uno de los pasos con que cuenta esta arquitectura. Después, en la sección 5.3 se presentan los resultados obtenidos al emplear dicha arquitectura utilizando el mismo corpus empleado para evaluar el método anterior. Asimismo, en la sección 5.4 se detallan los resultados

generados al utilizar la nueva arquitectura pero ahora sobre un conjunto de noticias de periódicos referentes al tema de Fútbol. Finalmente, en la sección 5.5 se concluye con una discusión sobre los resultados arrojados al usar ambos dominios.

5.1 Introducción

De acuerdo con el método explicado en el capítulo 4, este nuevo método se basa en los mismos objetivos que el anterior, ya que es no-supervisado, usa sólo información léxica, es fácil de llevarse a otro dominio y sigue las cuatro fases de los métodos basados en el enfoque no-supervisado (véase la sección 3.3). Sin embargo, lo que se busca con este nuevo método es descubrir la mayor cantidad posible de patrones léxicos de extracción, para lograrlo es necesario tener la mayor cantidad de información disponible. Es por ello que en esta nueva arquitectura ya no se aplica una técnica de minería de texto durante la fase “Descubrimiento de Patrones de Extracción”, sino un proceso de generación de ventanas, de igual manera durante la fase “Tipificación de Patrones” se realizará una tarea manual que consistirá en etiquetar sólo un pequeño conjunto de los ejemplos de entrada. Este conjunto de ejemplos serán a diferencia del método anterior patrones léxicos y no instancias de dichos patrones. Es importante señalar que en este método cada patrón léxico sólo podrá extraer información de un solo tipo a diferencia del método anterior, en el que cada patrón podría extraer información de distintos tipos. Además, como una de las características de este método es hacer mucho más fácil la tarea de llevarlo a otro dominio, se tiene como objetivo aplicarlo sobre dos dominios. Por tal motivo, se eligieron los dominios de: desastres naturales y fútbol. En el dominio de fútbol se desea extraer información como: nombre del equipo, marcadores, lugar del evento, entre otros.

5.2 Método propuesto

El método basado en patrones de extracción de tamaño fijo que se presenta en este capítulo, está construido siguiendo la arquitectura de los métodos basados en el enfoque no-supervisado (véase la figura 3.4). Esta arquitectura consta de cuatros fases: *Pre-procesamiento, Descubrimiento de Patrones de Extracción, Tipificación de Patrones* y *Generación de Plantillas.*

5.2.1 Fase 1: Pre-procesamiento

El objetivo de esta fase es generar un conjunto de oraciones organizadas por tipo (cantidades, fechas y nombres), a partir de una colección de documentos no etiquetados.

La forma en que se realizó esta fase fue la misma que se explica en la sección 4.2.1 del método descrito en el capítulo 4.

5.2.2 Fase 2: Descubrimiento de Patrones de Extracción

El objetivo de esta fase es descubrir un conjunto de patrones léxicos a partir del conjunto de oraciones organizadas por tipo resultantes de la fase anterior.

Para poder alcanzar el objetivo es necesario realizar un proceso de generación de ventanas. A continuación, se describe lo que se debe realizar en esta fase:

Tomar cada una de las oraciones sin palabras vacías y enseguida, tomar los contextos de un tamaño específico. Este tamaño también se le denomina

tamaño de ventana ($TamVen$). Es decir, se toma cada oración, de tal forma que a partir del elemento a extraer (por ejemplo, “@CANTIDAD@”, “@FECHA@” y “@NOMBRE@”) se realicé el corte de la oración tantas palabras a la izquierda y a la derecha de acuerdo al tamaño establecido en el tamaño de ventana ($TamVen$).

El resultado de realizar la actividad anterior es un conjunto de segmentos de textos organizados en cantidades, fechas y nombres de un tamaño fijo (también denominadas ventanas). Estas ventanas servirán como patrones léxicos de extracción, sin embargo, aún no se consideran útiles ya que no tienen identificado el tipo de información que extraerán.

Cabe mencionar que en este trabajo se emplearon los tamaños de ventana dos y tres, se eligieron estos valores porque se requiere de patrones generales para cumplir con su tarea, sin que sean demasiado pequeños en su estructura. Por ejemplo, en el método explicado en el capítulo 4, los patrones léxicos descubiertos mediante SFM tenían un tamaño que iba de dos a ocho. En cambio, con este método, es posible tener patrones léxicos con un tamaño menor, ya que su tamaño está relacionado con el tamaño de ventana dado.

Un ejemplo que muestre lo que se realiza en esta fase se describe a continuación.

Primeramente, consideremos las oraciones mostradas en la figura 5.1. Posteriormente, a éstas se les quitarán las palabras vacías (véase la figura 5.2).

1. Los incendios han ocasionado daños y @CANTIDAD@ bomberos sufrieron quemaduras cuando combatían las llamas.
2. Las inundaciones que han matado por lo menos @CANTIDAD@ personas han aumentado.
3. Las lluvias que han azotado el centro alcanza los @CANTIDAD@ muertos hasta el momento.
4. Hasta el momento existen @CANTIDAD@ muertos en la región.
5. El sismo acabo con la vida de @CANTIDAD@ personas.

Figura 5.1. Oraciones.

1. incendios ocasionado daños @CANTIDAD@ bomberos sufrieron quemaduras combatían llamas
2. inundaciones matado @CANTIDAD@ personas aumentado
3. lluvias azotado centro alcanza @CANTIDAD@ muertos momento
4. momento existen @CANTIDAD@ muertos región
5. sismo acabo vida @CANTIDAD@ personas

Figura 5.2. Oraciones sin palabras vacías.

Después, se hará un corte a cada una de las oraciones de la figura 5.2, para ello es necesario considerar un tamaño de ventana. Por ejemplo, si el tamaño de ventana elegido fuera dos, el conjunto de patrones léxicos resultantes sería el que se muestra en la figura 5.3.

1. ocasionado daños @CANTIDAD@ bomberos sufrieron
2. inundaciones matado @CANTIDAD@ personas aumentado
3. centro alcanza @CANTIDAD@ muertos momento
4. momento existen @CANTIDAD@ muertos región
5. acabo vida @CANTIDAD@ personas

Figura 5.3. Patrones léxicos.

Como se puede apreciar en la figura 5.3, lo único que se realizó fue cortar cada una de las oraciones, de tal forma que a partir de la etiqueta @CANTIDAD@ se dejaron dos palabras a la izquierda y dos a la derecha.

5.2.3 Fase 3: Tipificación de Patrones

El objetivo de esta fase es generar un conjunto de patrones léxicos específicos para un dominio a partir de un conjunto de patrones léxicos descubiertos en la fase anterior.

En la figura 5.4, se muestra la arquitectura que se siguió en esta fase para lograr el objetivo. Dicha arquitectura consta de tres actividades: Agrupamiento, Tipificación de Centroides y Ordenamiento de los Patrones.

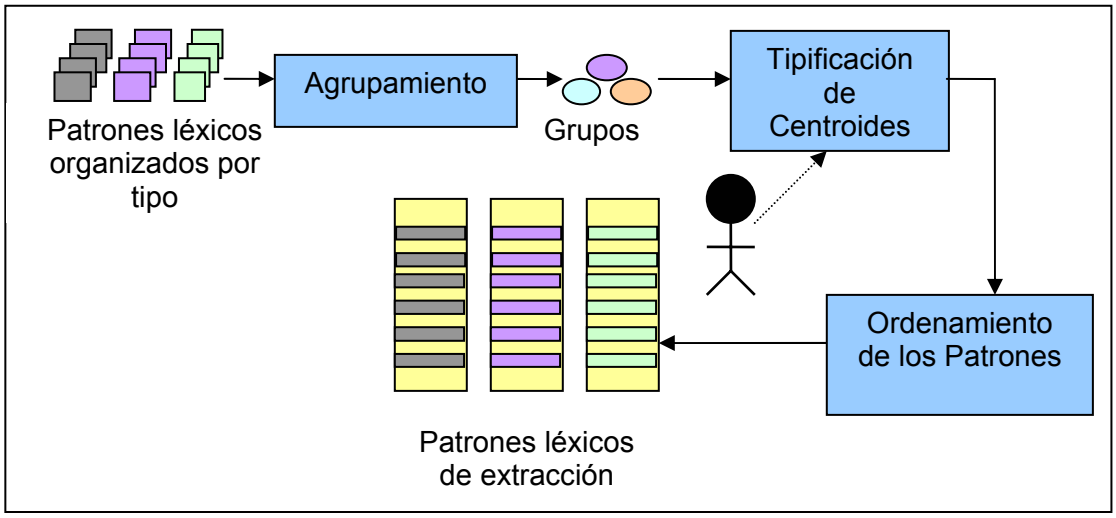

Figura 5.4. Arquitectura de la fase de Tipificación.

5.2.3.1 Agrupamiento

El objetivo de este paso es generar grupos de patrones léxicos a partir del conjunto de patrones léxicos resultantes de la fase 2.

Esta actividad es importante porque sirve como medio para categorizar los patrones léxicos descubiertos en la fase anterior. Como se recordará, dichos patrones están preclasificados en cantidades, fechas y nombres, pero aún no tienen asignado el tipo de información a extraer. Por tal motivo, no es posible utilizarlos para extraer información.

Para llevar a cabo esta actividad se deben formar grupos de patrones léxicos mediante un algoritmo de agrupamiento. Los patrones léxicos son los resultantes de la fase 2 y el algoritmo utilizado es el mismo que se empleó en el método anterior (algoritmo estrella).

El resultado de esta actividad es un conjunto de grupos de patrones con características similares entre ellos.

Esta tarea se lleva a cabo de la misma manera que como se explicó en la sección 4.2.3.2. Sin embargo, a diferencia del método anterior, en este caso se están agrupando patrones y no instancias.

5.2.3.2. Tipificación de Centroides

El objetivo de este paso es asignar el tipo de información que extraerán cada uno de los patrones léxicos a partir de los grupos obtenidos del paso anterior.

La forma en que se realizó esta tarea, fue la misma que se explica en la sección 4.2.3.3 del método descrito en el capítulo 4. Sin embargo, los centroides en este caso son patrones y no instancias. Además, aparte del centroide, también se proporcionó la frase asociada a éste, con el fin de facilitar el etiquetado.

5.2.3.3. Ordenamiento de los Patrones

El objetivo de este paso es generar un conjunto de patrones léxicos de extracción ordenados de acuerdo al tipo de información que extraerán.

Es importante mencionar que en este método, cada patrón sólo podrá extraer información de un sólo tipo. Esto difiere con el método expuesto en el capítulo 4, en el que se podía emplear un mismo patrón para extraer información de distintos tipos.

Un ejemplo que muestre cual sería el resultado de esta etapa se presenta en la tabla 5.1. En esta tabla se presenta un conjunto de patrones léxicos para extraer información sobre el número de muertos y el número de viviendas afectadas. Como se puede apreciar, en este caso los patrones no tienen un porcentaje asignado. Esto se debe a que los patrones extraerán información de un solo tipo.

CANTIDADES	
Tipo	**Patrón léxico**
Número de muertos	LLUVIAS DEJAN @CANTIDAD@ MUERTOS
	ENORMES DAÑOS @CANTIDAD@ MUERTOS PADECE
	DEJADO @CANTIDAD@ MUERTOS CERCA
	@CANTIDAD@ PERSONAS MUERTAS NUMEROSOS
Número de viviendas afectadas	DAÑÓ @CANTIDAD@ CASAS VECINOS
	EVACUAR @CANTIDAD@ CASAS QUEDARON
	TAPIADOS @CANTIDAD@ VIVIENDAS AFECTADAS

Tabla 5.1. Patrones léxicos

5.2.4 Fase 4: Generación de Plantillas

El objetivo de esta fase es generar un registro o plantilla por cada documento de entrada. Cabe mencionar que la información a extraer será establecida desde el inicio del proceso. Dicha información dependerá de los datos que sean de interés sobre el dominio a utilizar.

Para lograr el objetivo de esta fase es necesario aplicar los patrones léxicos de extracción resultantes de la fase “Tipificación de patrones”. Dichos patrones se aplicarán exclusivamente sobre textos que pertenezcan al mismo dominio al que pertenece el corpus utilizado en el método de extracción.

El resultado de esta fase será una plantilla por cada documento de entrada. Cada plantilla contendrá la información que se desea obtener de cada texto dado.

5.3 Resultados Experimentales en el dominio de Desastres Naturales

En esta sección se muestran los resultados obtenidos al aplicar este método sobre documentos que pertenecían al tema de desastres naturales. Al igual que el método anterior, los resultados de la evaluación se compararon con los obtenidos por TOPO.

De acuerdo con la arquitectura de este método de extracción (véase la figura 3.4), ésta consta de cuatro fases, de tal forma que para realizar los experimentos se emplearon para la fase 1, 500 noticias de periódicos sobre desastres naturales. De igual manera, en la fase 3 se utilizaron las mismas 500 noticias de la fase 1. Por último, en la fase 4 se realizó el proceso de evaluación utilizando 50 noticias de periódicos del mismo dominio.

El primer experimento que se llevó a cabo tenía como objetivo probar el nuevo método de extracción no-supervisado propuesto en un dominio sobre desastres naturales.

A continuación, se detallan los resultados obtenidos durante la realización de este experimento.

Durante la fase "*Descubrimiento de Patrones de Extracción*" se obtuvieron un conjunto de patrones léxicos (véase la tabla 5.2). En esta fase se utilizó un tamaño de ventana ($tamVen$) igual a 3.

Clases	**No. total de patrones léxicos**
CANTIDADES	3907
FECHAS	197
NOMBRES	5524

Tabla 5.2. Número total de patrones léxicos.

Durante la fase "*Tipificación de Patrones*" se utilizaron los patrones léxicos descubiertos en la fase "*Descubrimiento de Patrones de extracción*". Como se explicó en la sección 5.2.3.1, dichos patrones léxicos se agruparon mediante el algoritmo de agrupamiento estrella, para lo cual se consideraron los umbrales, μ y $\mu+\sigma$ (véanse las ecuaciones 4.1 y 4.2).

Asimismo, como se explicó en la sección 5.2.3.2, la siguiente actividad fue etiquetar los centroides de los grupos formados. En la tabla 5.3, se muestra la cantidad de centroides etiquetados en este experimento.

Clases	**No. total de centroides etiquetados**	
	μ	**μ+σ**
CANTIDADES	108	160
FECHAS	15	16
NOMBRES	194	218

Tabla 5.3. Centroides etiquetados manualmente.

De la tabla 5.3 se destaca lo siguiente:

- Se etiquetaron 317 centroides cuando se empleó el umbral μ y 394 cuando se utilizó el umbral μ+σ. En comparación con el método explicado en el capítulo 4, con este método se etiquetaron 62 centroides más que con el anterior, utilizando el $umbral = \mu + \sigma$.

Es importante destacar que la cantidad total de centroides etiquetados en este experimento es mucho menor en comparación con las aproximadamente 6600 oraciones que TOPO hubiera etiquetado.

Categorías	No. total de patrones	
	μ	**μ+σ**
CANTIDADES		
EVE_MAGNITUD	0	402
PER_MUERTAS	898	858
PER_HERIDAS	60	116
PER_DESAPARECIDAS	5	4
PER_DAMNIFICADAS	7	94
PER_AFECTADAS	145	216
VIV_DESTRUIDAS	26	132
VIV_AFECTADAS	191	3
INF_HECTAREAS	285	213
INF_ECONOMICA	0	3
FECHAS		
EVE_FECHA	77	60
NOMBRES		
EVE_LUGAR	2881	2213

Tabla 5.4. Cantidad total de patrones léxicos.

Con base en lo anterior, en la tabla 5.4 se muestra la cantidad de patrones léxicos de extracción obtenidos como resultado de la fase 3. De esta tabla se puede destacar lo siguiente:

- Se incrementó la cantidad de patrones léxicos de extracción. Anteriormente, con el método expuesto en el capítulo 4 se obtuvieron un máximo de 599 patrones léxicos, en cambio con este nuevo método se obtuvieron 4575 patrones empleando el $umbral = \mu$ y 4314 utilizando el $umbral = \mu + \sigma$.

Finalmente, durante la fase "*Generación de Plantillas*" se realizó el proceso de extracción de información cuya plantilla de salida fue la que se muestra en la figura 4.9, para lograrlo se aplicaron los patrones léxicos de extracción resultantes a 50 noticias de desastres naturales. Estas noticias fueron las mismas con que se evaluó el TOPO. Al final, se evaluó el método de extracción propuesto con base en las medidas de precisión, recuerdo y medida-F (véase la tabla 5.5).

Experimento	**Evaluaciones**	**Precisión**	**Recuerdo**	**Medida-F**
	TOPO	0.809	0.885	0.845
No. 1	Método propuesto (μ)	0.741	0.743	0.742
	Método propuesto (μ+σ)	0.741	0.744	0.743

Tabla 5.5. Resultados de la evaluación del método de extracción con un tamaño de ventana igual a 3.

Evaluaciones	**Precisión**	**Recuerdo**	**Medida-F**
TOPO	0.809	0.885	0.845
No.1	0.715	0.751	0.733
No.2	0.646	0.767	0.701

Tabla 5.6. Resultados de la evaluación del método basado en patrones de extracción de tamaño variable.

De acuerdo con los resultados de este experimento se puede destacar lo siguiente:

- Se mejoraron los resultados en comparación con los obtenidos con el método explicado en el capítulo 4 (véase la tabla 5.6).
- Se comprobó que al haber una mayor cantidad de patrones léxicos, los resultados de la evaluación se incrementan. Estos resultados van de un 73.3% de medida-F que fue el resultado más alto en el método anterior a un 74.3% obtenido como resultado máximo en este método.

Sin embargo, aunque los resultados fueron mejores que los obtenidos con el método anterior, se observó que los patrones eran aún muy particulares. Esto motivó que se llevara a cabo otro experimento, en el que se disminuyera el tamaño de los patrones léxicos, con el fin de que éstos fueran aún más generales.

El objetivo del segundo experimento fue probar el nuevo método de extracción no-supervisado propuesto en un dominio sobre desastres naturales, disminuyendo el tamaño de los patrones.

Para este experimento, se empleó un tamaño de ventana igual a 2 para la fase "*Descubrimiento de Patrones de Extracción*". En la tabla 5.7, se muestra la cantidad de patrones léxicos resultantes de esta fase.

Clases	**No. total de patrones léxicos**
CANTIDADES	3706
FECHAS	130
NOMBRES	5102

Tabla 5.7. Número total de patrones léxicos.

Durante la fase "*Tipificación de Patrones*" se utilizaron los patrones léxicos descubiertos en la fase "*Descubrimiento de Patrones de Extracción*". Como se explicó en la sección 5.2.3.1, dichos patrones léxicos se agruparon mediante el algoritmo de agrupamiento estrella, para lo cual se consideró el $umbral = \mu + \sigma$, ya que fue éste el que mejores resultados arrojó de acuerdo con los resultados mostrados en la tabla 5.5.

Asimismo, la siguiente actividad después de agrupar los patrones fue etiquetar los centroides de los grupos formados. En la tabla 5.8, se muestra la cantidad de centroides etiquetados en este experimento.

Clases	**No. total de centroides etiquetados**
	μ+σ
CANTIDADES	145
FECHAS	9
NOMBRES	248

Tabla 5.8. Centroides etiquetados manualmente.

De la tabla 5.8 se destaca lo siguiente:

- La cantidad de centroides etiquetados aumentó de 394 a 402 con el segundo experimento.

Con base en lo anterior, en la tabla 5.9 se muestra la cantidad de patrones léxicos de extracción obtenidos como resultado de la fase 3.

Categorías	No. total de patrones
	μ+σ
CANTIDADES	
EVE_MAGNITUD	334
PER_MUERTAS	673
PER_HERIDAS	17
PER_DESAPARECIDAS	6
PER_DAMNIFICADAS	3
PER_AFECTADAS	94
VIV_DESTRUIDAS	102
VIV_AFECTADAS	106
INF_HECTAREAS	210
INF_ECONOMICA	3
FECHAS	
EVE_FECHA	44
NOMBRES	
EVE_LUGAR	2308

Tabla 5.9. Cantidad total de patrones léxicos.

Finalmente, durante la fase "*Generación de Plantillas*" se realizó el proceso de extracción de información, para lograrlo se aplicaron los patrones léxicos de extracción resultantes a las mismas 50 noticias utilizadas en el primer experimento. Al final, se evaluó el método de extracción propuesto con base en las medidas de precisión, recuerdo y medida-F. Los resultados obtenidos se muestran en la tabla 5.10.

Experimento	Evaluaciones	Precisión	Recuerdo	Medida-F
	TOPO	0.809	0.885	0.845
No. 1	Método propuesto (μ)	0.741	0.743	0.742
	Método propuesto (μ+σ)	0.741	0.744	0.743
No. 2	Método propuesto (μ+σ)	0.736	0.773	0.754

Tabla 5.10. Resultados de la evaluación.

De los resultados generados del proceso de evaluación se puede destacar lo siguiente:

- El recuerdo aumentó de un 74.4% a un 77.3%.
- La precisión se redujo de un 74.1% a un 73.6%.
- Se mejoraron los resultados con respecto al primer experimento de un 74.3% a un 75.4% de medida-F.
- El método de extracción de información propuesto está a un 9.1% de medida-F por debajo del TOPO.

Es importante señalar que los métodos que se basan en el enfoque supervisado, por sus características, son más precisos que los métodos que utilizan el enfoque no-supervisado (véanse las secciones 3.2 y 3.3). Sin embargo, los métodos supervisados requieren de un mayor esfuerzo, principalmente manual, en comparación con los métodos no-supervisados.

Por lo anterior, los resultados obtenidos con el método propuesto son satisfactorios. Aunque, el TOPO supera a este método en un 9.1% de medida-F, el método propuesto es más portable que el TOPO. Esto se debe en gran medida a la disminución del trabajo manual realizado.

5.4 Resultados Experimentales en el Dominio de Fútbol

En esta sección se presentan los resultados obtenidos al evaluar el “Método Basado en Patrones de Extracción de Tamaño Fijo” empleando noticias de periódicos pertenecientes al tema de fútbol. El tipo de información que se

maneja en este dominio se muestra en la tabla 5.11. El objetivo de evaluar con otro dominio distinto es comprobar la portabilidad de este método de extracción.

CANTIDADES	
MARCADOR	Número de goles obtenidos por ambos equipos al final del partido de fútbol.
FECHAS	
EVE_FECHA	Fecha en que ocurrió el partido de fútbol.
NOMBRES	
LUGAR	Estadio donde se llevó a cabo el partido de fútbol.
EQUIPOS	Nombre de los equipos que jugaron en el partido de fútbol.
E_GANADOR	Nombre del equipo que resultó vencedor en el partido de fútbol.

Tabla 5.11. Tipos de información para el dominio de Fútbol.

De acuerdo con la arquitectura de este método de extracción (véase la figura 3.4), ésta consta de cuatro fases, de tal forma que para realizar los experimentos se emplearon para la fase 1, 182 noticias de periódicos sobre fútbol. De igual manera, en la fase 3 se utilizaron las mismas 182 noticias de la fase 1. Por último, en la fase 4 se realizó el proceso de evaluación utilizando 20 noticias de periódicos del mismo dominio.

El primer experimento que se llevó a cabo utilizando noticias sobre fútbol, tenía como objetivo probar el nuevo método de extracción no-supervisado propuesto en un nuevo dominio.

A continuación, se detallan los resultados obtenidos durante la realización de este experimento.

Durante la fase "*Descubrimiento de Patrones de Extracción*" se obtuvieron un conjunto de patrones léxicos (véase la tabla 5.12). En esta fase se utilizó un tamaño de ventana ($tamVen$) igual a 3.

Clases	No. total de patrones léxicos
CANTIDADES	458
FECHAS	201
NOMBRES	6393

Tabla 5.12. Número total de patrones léxicos.

Durante la fase "*Tipificación de Patrones*" se utilizaron los patrones léxicos descubiertos durante la fase "*Descubrimiento de patrones de extracción*". Como se explicó en la sección 5.2.3.1, dichos patrones léxicos se agruparon mediante el algoritmo de agrupamiento estrella, para lo cual se consideraron los umbrales μ y $\mu+\sigma$ (véanse las ecuaciones 4.1 y 4.2).

Asimismo, como se explicó en la sección 5.2.3.2, la siguiente actividad fue etiquetar los centroides de los grupos formados. En la tabla 5.13, se muestra la cantidad de centroides etiquetados en este experimento.

Clases	No. total de centroides etiquetados	
	μ	**μ+σ**
CANTIDADES	0	0
FECHAS	14	14
NOMBRES	181	188

Tabla 5.13. Centroides etiquetados manualmente.

De la tabla 5.13 se destaca lo siguiente:

- Se etiquetaron 195 centroides cuando se empleo el umbral μ y 202 cuando se utilizó el umbral μ+σ.

- No se etiquetó ningún centroide para la clase CANTIDADES. Esto se debió a que la única cantidad que se identificó en los textos de entrada fueron “marcadores del partido de fútbol”. Por tal motivo, no fue necesario realizar un proceso de tipificación para este caso.

Categorías	No. total de patrones	
	μ	μ+σ
CANTIDADES		
MARCADOR	458	458
FECHAS		
EVE_FECHA	11	7
NOMBRES		
LUGAR	12	529
EQUIPOS	1936	1213
E_GANADOR	46	226

Tabla 5.14. Cantidad total de patrones léxicos.

Con base en lo anterior, en la tabla 5.14 se muestra la cantidad de patrones léxicos de extracción obtenidos como resultado de la fase 3. De esta tabla se puede destacar lo siguiente:

- Se descubrieron 2463 patrones léxicos empleando un $umbral = \mu$ y 2433 utilizando un $umbral = \mu + \sigma$.

INFORMACIÓN DEL PARTIDO	
Fecha	
Lugar	
Equipos	
Equipo ganador	
Marcador final	

Figura 5.5. Plantilla de extracción para el dominio de Fútbol.

Finalmente, durante la fase “*Generación de Plantillas*” se realizó el proceso de extracción de información, cuya plantilla de salida fue la que se muestra en la figura 5.5, para lograrlo se aplicaron los patrones léxicos de extracción resultantes a 20 noticias de fútbol. Al final, se evaluó el método de extracción

propuesto con base en las medidas de precisión, recuerdo y medida-F (véase la tabla 5.15).

Experimento	Evaluaciones	Precisión	Recuerdo	Medida-F
No. 1	Método propuesto (μ)	0.466	0.494	0.480
	Método propuesto (μ+σ)	0.477	0.505	0.490

Tabla 5.15. Resultados de la evaluación.

De acuerdo con los resultados de este experimento se puede destacar lo siguiente:

- El método propuesto en este capítulo obtiene mejores resultados cuando se utiliza el $umbral = \mu + \sigma$.

Sin embargo, en este experimento se observó que los patrones léxicos eran muy particulares debido a su tamaño. Esto motivó a que se llevara a cabo otro experimento, en el que se disminuyera el tamaño de dichos patrones. El objetivo de esto era hacer más generales los patrones.

El objetivo del segundo experimento fue probar el nuevo método de extracción no-supervisado propuesto en un dominio sobre fútbol, disminuyendo el tamaño de los patrones.

Para este experimento, se empleó un tamaño de ventana igual a 2 para la fase "*Descubrimiento de Patrones de Extracción*". En la tabla 5.16, se muestra la cantidad de patrones léxicos resultantes de esta fase.

Clases	No. total de patrones léxicos
CANTIDADES	449
FECHAS	180
NOMBRES	6176

Tabla 5.16. Número total de patrones léxicos.

Durante la fase "*Tipificación de Patrones*" se utilizaron los patrones léxicos descubiertos en la fase "*Descubrimiento de Patrones de Extracción*". Como se explicó en la sección 5.2.3.1, dichos patrones léxicos se agruparon mediante el algoritmo de agrupamiento estrella. Para lo cual se consideraron los umbrales μ y $\mu + \sigma$ (véanse las ecuaciones 4.1 y 4.2).

Asimismo, la siguiente actividad después de agrupar los patrones fue etiquetar los centroides de los grupos formados. En la tabla 5.17, se muestra la cantidad de centroides etiquetados en este experimento.

Clases	No. total de centroides etiquetados	
	μ	μ+σ
CANTIDADES	0	0
FECHAS	7	7
NOMBRES	245	245

Tabla 5.17. Centroides etiquetados manualmente.

De la tabla 5.17 se destaca lo siguiente:

- La cantidad de centroides etiquetados aumentó con el segundo experimento de 195 a 252 utilizando ambos umbrales.

Con base en lo anterior, en la tabla 5.18 se muestra la cantidad de patrones léxicos de extracción obtenidos como resultado de la fase 3.

Categorías	No. total de patrones	
	μ	μ+σ
CANTIDADES		
MARCADOR	449	449
FECHAS		
EVE_FECHA	3	3
NOMBRES		
LUGAR	254	242
EQUIPOS	726	715
E_GANADOR	158	127

Tabla 5.18. Cantidad total de patrones léxicos.

Finalmente, durante la fase "*Generación de plantillas*", se realizó el proceso de extracción de información sobre las mismas 20 noticias del primer experimento. Al final, se evaluó el método de extracción propuesto con base en las medidas de precisión, recuerdo y medida-F (véase la tabla 5.19).

Experimento	Evaluaciones	Precisión	Recuerdo	Medida-F
No. 1	Método propuesto (μ)	0.466	0.494	0.480
	Método propuesto (μ+σ)	0.477	0.505	0.490
No. 2	Método propuesto (μ)	0.514	0.542	0.527
	Método propuesto (μ+σ)	0.524	0.552	0.538

Tabla 5.19. Resultados de la evaluación.

De los resultados generados del proceso de evaluación se puede destacar lo siguiente:

- Los resultados en este experimento fueron mejores en comparación con los resultados del experimento anterior.
- Se mejoró el recuerdo, lo cual significa que de la información que se debía extraer se pudo obtener con un máximo del 55.2%.

- Los resultados en precisión se mejoraron, sin embargo, sólo se pudo lograr un 52.4%. Esto se debió a que mucha de la información recopilada estaba incorrecta y por consiguiente, se redujo la precisión.

5.5 Discusión

De acuerdo con los resultados mostrados en este capítulo, se puede resaltar que este método logra obtener una mayor cantidad de patrones léxicos útiles para el proceso de extracción de información. Además, para el caso del dominio de desastres naturales, mejora los resultados de evaluación con respecto a los obtenidos con el método explicado en el capítulo 4. Esto se debe al descubrimiento de una mayor cantidad de patrones léxicos de extracción útiles. Sin embargo, aunque no se obtuvieron mejores resultados que los obtenidos por el TOPO, es posible definir que los resultados generados por este método son adecuados. Esto se debe a que ambos métodos emplean técnicas distintas para cumplir con la tarea de extracción y además, se basan en enfoques distintos. Es importante señalar que este método requiere de un trabajo manual mucho menor en comparación con el TOPO, lo cual se considera una ventaja para facilitar su traslado a otro dominio.

Por otro lado, en el dominio de fútbol los resultados máximos fueron del 53.8% de medida-F. Esto se debió a la cantidad de noticias del corpus de fútbol, el cual está integrado por sólo 202 textos. En cambio, el corpus utilizado para el dominio de desastres naturales está integrado por 550 noticias. Además, se notó que existe en el dominio de fútbol una enorme cantidad de expresiones utilizadas para extraer información, por lo tanto, es necesario tener más noticias para probar el método propuesto en este

dominio específico. Este hecho ocasionó que se alcanzará como máximo un 52.4% en precisión y un 55.2% en recuerdo.

De acuerdo con [Cardino G. et al., 1997] la portabilidad es la facilidad con la cual un sistema puede ser transferido de un medio ambiente a otro. Además, este concepto esta asociado a la dependencia que tiene el sistema de elementos funcionales externos, como por ejemplo, patrones lingüísticos, palabras semilla, entre otros. De acuerdo con [Cardino G. et al., 1997] para lograr una alta portabilidad es necesario que el número de elementos funcionales dependientes del sistema sea bajo. El método propuesto solo necesita la modificación de dos cosas para ser llevado a otro dominio: la lista de palabras vacías y las expresiones regulares utilizadas para identificar cantidades, fechas y nombres.

Por lo anterior, se puede determinar que el método propuesto tiene una alta portabilidad, ya que pudo ser llevado a otro dominio sin requerir de grandes modificaciones, y además, no necesita de elementos preestablecidos, como por ejemplo, un conjunto de patrones léxicos predefinidos.

Capítulo 6

Conclusiones

6.1 Sumario

En este trabajo de tesis se propusieron dos métodos de extracción de información no-supervisados. El objetivo de ambos métodos es identificar descripciones de eventos de textos en lenguaje natural y por consiguiente, extraer la información relacionada a dichos eventos.

Ambos métodos propuestos tienen como característica principal que son más portables en comparación con los métodos existentes. Sin embargo, sólo se comprobó la portabilidad del segundo método mediante su evaluación en dos dominios (Desastres Naturales y Fútbol).

El primer método descubre patrones léxicos de tamaño variable mediante una técnica de minería de texto. Posteriormente, genera un conjunto de instancias a partir de los patrones descubiertos. Dichas instancias se agrupan y posteriormente, se tipifica sólo una pequeña porción de éstas. El resultado es un conjunto de patrones léxicos de extracción que podrán

extraer información de tipos distintos. Por tal motivo, dichos patrones están ponderados.

El segundo método se caracteriza por no emplear una técnica de minería de texto. Sin embargo, se sustituyó esta característica por un proceso de generación de ventanas, donde el resultado de éste es un conjunto de patrones léxicos de tamaño fijo. Dichos patrones en cantidad son más en comparación con los descubiertos con el primer método. Después del proceso de generación de ventanas, los patrones léxicos resultantes se agrupan y posteriormente, se tipifica sólo una pequeña porción de éstos. Una vez tipificados los patrones, lo siguiente es utilizarlos como conocimiento base para el proceso de extracción de información.

Los resultados alcanzados demostraron que el uso del segundo método propuesto (véase capítulo 5) en el dominio de desastres naturales, ayuda a descubrir una mayor cantidad de patrones léxicos de extracción, lo cual ocasionó que los resultados fueran mejores que con el método anterior. Sin embargo, estos resultados están a un 9.1% por debajo del TOPO. Aunque, es posible definir que los resultados generados por este método son adecuados. Esto se debe a que tanto el TOPO como el método propuesto emplean técnicas distintas para cumplir con la tarea de extracción y además, se basan en enfoques distintos. Es importante mencionar que este método requiere de un trabajo manual mucho menor en comparación con el TOPO, lo cual se considera una ventaja para facilitar su traslado a otro dominio.

Por otro lado, al evaluar el segundo método propuesto en el dominio de fútbol, los resultados fueron adecuados considerando que lo único que se quería probar era la portabilidad del método. Aunque, cabe mencionar que debido a la enorme cantidad de expresiones utilizadas para extraer información en este dominio era necesario un corpus de entrada mucho

mayor al que se tenía, de tal forma que se obtuviera una mayor cantidad de patrones léxicos a la que se descubrió. El objetivo de esto era tener patrones léxicos que permitieran extraer información expresada en múltiples formas.

6.2 Conclusiones

Para el "Método Basado en Patrones de Extracción de Tamaño Variable" no se descubrió la cantidad suficiente de patrones de extracción, debido a que se perdía mucha información durante la fase "*Descubrimiento de Patrones*". Este hecho ocasionó que en las fases siguientes del método se tuviera poca información.

Ambos métodos requieren que el tamaño del corpus empleado sea suficientemente grande para obtener la mayor cantidad de patrones léxicos de extracción. Lo anterior se debe a que al tener una mayor cantidad de documentos se obtiene una mayor cantidad de formas en que se expresa la información en dichos documentos, originando que se descubran una mayor diversidad de patrones, lo cual ocasiona que se alcancen mejores resultados de evaluación.

Finalmente, se concluyo que ambos métodos propuestos son útiles en el proceso de extracción de información. Además, son más portables en comparación con otros métodos.

6.3 Trabajo Futuro

A continuación, se enlistan las actividades futuras que serían de utilidad para mejorar los métodos expuestos en esta tesis.

1. Mejorar la tarea de reconocimiento de entidades nombradas. Esta tarea actualmente cuenta con muchas fallas, ya que ésta actúa como un detector de nombres, sin embargo, no identifica la clase semántica (por ejemplo, lugar, nombre de persona, organización, entre otros). Esto conlleva a que entidades no útiles para el método sean empleadas durante el proceso de extracción. Es por ello que sería conveniente utilizar un identificador de entidades nombradas más especializado. Este identificador ayudaría a obtener únicamente la información que se desea, lo cual permitiría alcanzar una mayor precisión. Cabe mencionar, que el identificador de entidades nombradas utilizado en esta tesis fue el empleado en [Téllez A., 2005].

2. Buscar una alternativa que ayude a encontrar la cantidad óptima de textos a utilizar, de tal forma que en dichos textos se maneje la mayoría de las formas en que se expresa la información a extraer, para lograrlo es necesario buscar una manera donde se elimine desde el inicio del proceso los textos irrelevantes, es decir, aquellos textos que por su contenido carecen de importancia para la tarea, logrando con ello mantener sólo los textos útiles para el proceso.

3. Aplicar los métodos propuestos sobre otros dominios en distintos idiomas, de tal forma que se compruebe su independencia al idioma, debido a que se emplean únicamente elementos léxicos.

Índice de Figuras

Índice de Tablas

Bibliografía

[Ahonen-Myka H.,1999] Ahonen-Myka H. “Knowledge Discovery in Documents by Extracting Frequent Word Sequences”. Library trends, vol. 48, no. 1, pp. 160-181, 1999.

[Ahonen-Myka H.,2002] Ahonen H. “Discovery of Frequent Word Sequences in Text”. Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery. London, UK, pp. 180-189, 2002.

[Aslam J. et al.,1999] Aslam J., Pelekhov K. & Rus D. “A Practical Clustering Algorithm for Static and Dynamic Information Organization”. In Proceedings of the 1999 Symposium on Discrete Algorithms, pp. 208-217, 1999.

[Aslam J. et. al.,2000] Aslam J., Pelekhov K. & Rus D. “Information Organization Algorithms”. In Proceedings of the International Conference on Advances in Infrastructure for Electronic Business (SSGRR), Science, and Education on the Internet, 2000.

[Brin S., 1998] Brin S. “Extracting Patterns and Relations from the World-Wide Web”. 1998 Int'l Workshop on the Web and Databases (WebDB '98), pp.172-18, 1998.

[Califf M. & Mooney R., 2003] Califf M. & Mooney R. “Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction”. Journal of Machine Learning Research, 4:177–210, 2003.

[Callan J. & Mitamura T.,2002] Callan J. & Mitamura T. Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, Virginia, USA, pp. 532-537, 2002.

[Cardino G. et al., 1997] Cardino G., Baruchelli F. & Valerio A. “The Evaluation of Framework Reusability”. ACM SIGAPP Applied Computing Review, New York, USA, pp. 21-27, 1997.

[Chang C. & Lui S., 2001] Chang C. & Lui S. “IEPAD: Information Extraction Based on Pattern Discovery”. In Proceedings International WWW Conference (10), Hong-Kong, pp. 681-688, 2001.

[Chinchor N. & Sundheim B., 1993] Chinchor N. & Sundheim B. “MUC-5 Evaluation Metrics**”.** Proceedings of the 5th Conference on Message Understanding, Baltimore, Maryland, pp. 69-78, 1993.

[Chinchor N., 1998] Chinchor N. “MUC-7 Test Scores Introduction”. In Proceedings of the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[Ciravegna F., 2001] Ciravegna F. “Adaptive Information Extraction from Text by Rule Induction and Generalization”. In Proceedings 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, pp. 1251-1256, 2001.

[Cowie J. & Lehnert W., 1996] Cowie J. & Lehnert W. “Information Extraction”. Communications of the ACM. 39(1) :80-91,1996.

[Craven M. et al.,1998] Craven M., DiPasquo D., Freitag D., McCallum A., Mitchell T., Nigam K. & Slattery S. “Learning to Extract Symbolic Knowledge from the World Wide Web”. In Proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. 509-516, 1998.

[Feldman R. & Sanger J., 2007] Feldman R. & Sanger J. “The Text Mining Handbook”. Cambridge University Press, Nueva York, 2007.

[García R. et al.,2006] García R., Martínez F. & Carrasco A. “A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection”. International Conference on Computational Linguistics and Text Processing, CICLing-2006, Mexico City, Mexico, pp. 514-523, 2006.

[Grishman R., 1993] Grishman R. **“**Design of the MUC-6 Evaluation**”.** Proceedings of the 6th Conference on Message understanding, Columbia, Maryland, pp. 1-11, 1993.

[Han J. & Kamber M.,2001] Han J. & Kamber M. “Data Mining: Concepts and Techniques”. Elsevier, 2nd edition, 2001.

[Hernández J. et al., 2004] Hernández J., Rámirez M. & Ferri C. “Introducción a la Minería de Datos”. Prentice Hall, Pearson Educación, S.A., Madrid, 2004.

[Hoff G., 2002] G. Hoff. HomePageSearch. http://hpsearch.uni-trier.de/. University of Trier, 2002.

[Jain A.et al.,1999] Jain A., Murty M. & Flynn P. “Data Clustering: A Review”. ACM Computing Surveys, pp. 264-323, 1999.

[Kovács L. & Ahonen-Myka H.,2001] Kovács L. & Ahonen-Myka H. “Algorithm for Maximal Frequent Sequences in Document Clustering”. 3rd International Symposium of Hungarian Researchers on Computational Intelligence, pp. 89-94, 2001.

[Mackhoul J. et al., 1999] Makhoul J., Kubala F., Schwartz R. & Weischedel R. “Performance Measures for Information Extraction”. In Proceedings of DARPA Broadcast News Workshop,pp. 249-254, 1999.

[Morrison D., 1968] Morrison D. “PATRICIA-Practical Algorithm to Retrieve Information Coded in Alphanumeric”. Journal of ACM (JACM), pp. 514-534, 1968.

[Patward S. & Riloff E.,2006] Patwardham S. & Riloff E. “Learning Domain-Specific Information Extraction Patterns from the Web”. Proceedings of the ACL 2006 Workshop on Information Extraction Beyond the Document, pp. 66-73, 2006.

[Riloff E., 1993] Riloff E. “Automatically Constructing a Dictionary for Information Extraction Tasks”. 11th National Conference on Artificial Intelligence (AAAI), pp. 811-816, 1993.

[Riloff E., 1996] Riloff E. “Automatically Generating Extraction Patterns from Untagged Text”. 13th National Conference on Artificial Intelligence (AAAI), pp. 1044-1049, 1996.

[Riloff E. & Jones R., 1999] Riloff E. & Jones R. “Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping”. Sixteenth National Conference on Artificial Intelligence (AAAI), pp. 474-479, 1999.

[Riloff E. & Phillips W., 2004] Riloff E. & Phillips W. “An Introduction to the Sundance and AutoSlog Systems”. Technical Report UUCS-04-015, School of Computing, University of Utah, 2004.

[Riloff E. et al., 2002] Riloff E., Schafer C. & Yarowsky D. “Inducing Information Extraction Systems for New Languages via Cross-Language Projection”. In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), pp. 1-7, 2002.

[Riloff E. et al., 2005] Riloff E., Wiebe J. & Phillips W. “Exploiting Subjectivity Classification to Improve Information Extraction”. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05), pp. 1106-1111, 2005.

[Sabou M., 2005] Sabou M. “Learning Web Service Ontologies: an Automatic Extraction Method and its Evaluation”. In: Ontology Learning from text: Methods, Evaluation and Applications, IOS Press, Armsterdam, Nertherlands,pp. 125-139, 2005.

[Soderland S. et. al., 1995] Soderland S., Aronow D., Fisher D., Aseltine J. & Lehnert W. “Machine Learning of Text-Analysis Rules for Clinical Records”. Technical Report of the University of Massachussets, 1995.

[Sudo K. et al., 2001] Sudo K., Sekine S. & Grishman R. “Automatic Pattern Acquisition for Japanese Information Extraction”. First International Conference on Human Language Technology Research, pp. 1-7, 2001.

[Sundheim B., 1993] Sundheim B. **“**Overview of Results of the MUC-6 Evaluation**”**. Proceedings of the 6th Conference on Message understanding, Columbia, Maryland, pp. 13-31, 1993.

[Téllez A., 2005] Téllez A. “Extracción de Información con Algoritmos de Clasificación”. Tesis de Maestría en Ciencias Computacionales. Instituto Nacional de Astrofísica, Óptica y Electrónica. Puebla, México, 2005.

[Téllez A. et al., 2005] Téllez A, Montés M. & Villaseñor L. “A machine Learning Approach to Information Extraction”. International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005), pp. 527 - 536, 2005.

[Turmo J. et al., 2006] Turmo J., Ageno A. & Catalá N. “Adaptive Information Extraction”. Proceedings of the ACM Computing Surveys, Barcelona, Spain. 38 (2): 1-47, 2006.

[Van-Rijsbergen C. ,1979] Van-Rijsbergen C. “Information Retrieval”. Butterworths, London, England, 2nd edition, 1979.

[Weizenbaum J.,1966] Weizenbaum J. “ELIZA – A Computer Program for the Study of Natural Language Communications between Men and Machines”. Communications of the Association for Computing Machinery, 9: 36–45, 1966.

[Yangarber et al., 2000] Yangarber R, Grishman R, Tapanainen P. & Huttunen S. “Unsupervised Discovery of Scenario-level Patterns for Information Extraction”. 6th Applied Natural Language Processing Conference (ANLP), pp. 282-289, 2000.