SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Secretaría Académica, de Investigación e Innovación
Dirección de Posgrado, Investigación e Innovación

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

# Centro Nacional de Investigación y Desarrollo Tecnológico

## Subdirección Académica

## Departamento de Ciencias Computacionales

## TESIS DE MAESTRÍA EN CIENCIAS

### Extracción Automática de Hechos de Noticias de Desastres Naturales Escritas en Español

presentada por
**L.I. Vania Mydori Castillo Rendón**

como requisito para la obtención del grado de
**Maestra en Ciencias de la Computación**

Director de tesis
**Dr. Noé Alejandro Castro Sánchez**

**Cuernavaca, Morelos, México. Julio de 2016.**

Cuernavaca, Morelos a 20 de junio del 2016
OFICIO No. DCC/157/2016
**Asunto:** Aceptación de documento de tesis

**C. DR. GERARDO V. GUERRERO RAMÍREZ**
**SUBDIRECTOR ACADÉMICO**
**PRESENTE**

Por este conducto, los integrantes de Comité Tutorial del **Lic. Vania Mydori Castillo Rendón**, con número de control M14CE009, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado **"Extracción automática de hechos de noticias de desastres naturales escritas en español"** y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806

REVISOR 1

Dra. Alicia Martínez Rebollar
Doctora en Informática
7399055

REVISOR 2

Dr. Luis Gerardo Vela Valdés
Doctor En Ciencias En Ingeniería Electrónica
7980044

REVISOR 3

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la Computación
7820329

C.p. Lic. Guadalupe Garrido Rivera - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

AMR/lmz

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 23 de junio de 2016
OFICIO No. SAC/237/2016

**Asunto:** Autorización de impresión de tesis

**LIC. VANIA MYDORI CASTILLO RENDÓN**
**CANDIDATA AL GRADO DE MAESTRA EN CIENCIAS DE LA COMPUTACIÓN**
**P R E S E N T E**

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Extracción automática de hechos de noticias de desastres naturales escritas en español"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

**ATENTAMENTE**
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

**DR. GERARDO VICENTE GUERRERO RAMÍREZ**
**SUBDIRECTOR ACADÉMICO**

SEP TecNM
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

C.p. Lic. Guadalupe Garrido Rivera.- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

# Dedicatoria

Para mi padre Raúl Castillo Miranda y mi hermano Eduardo Castillo Rendón. Ellos quienes son mis impulsores en esta vida y que por ellos he cumplido con esta etapa. Quienes no sueltan mi mano por más complicado que de torne el momento.

Para mi novio, Salvador Montes Martínez quien con su apoyo deja en mí una huella imborrable en mí. Te amo y siempre te estaré infinitamente agradecida.

Mi familia, la familia Castillo Miranda y a la familia Montes Martínez que en ningún momento dejaron de creer en mí. Su apoyo incondicional me promovió a seguir adelante.

Mi director de tesis, Dr. Noé A. Castro, sé que en un momento me veía aislarme pero su insistencia, paciencia y sobre todo su confianza me ayudaron para que se pudiéramos terminar este trabajo.

**"Siempre que te pregunten si puedes hacer un trabajo,**
**contesta que sí y ponte enseguida a aprender cómo se hace"**
Franklin D. Roosevelt.

# Agradecimientos

A Dios, por permitirme vivir estos momentos de la vida. Por poner en mí camino a las personas indicadas.

Debo agradecer al Consejo Nacional de Ciencia y Tecnología (CONACYT) por su apoyo económico que me permitió solventarme durante el desarrollo de esta investigación.

Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), por la oportunidad de brindarme una estancia en su institución para realizar la Maestría en Ciencias de la Computación. Así como al personal académico y administrativo quienes con su trabajo ayudan al desarrollo de nuestras actividades.

En especial al Dr. Noé Alejandro Castro quien me dio la oportunidad de trabajar en conjunto para poder realizar esta investigación. Por su orientación día a día para que el trabajo pudiera finalizarse. Además de brindarme no solo su confianza, sino también su amistad, ya que fue de gran apoyo en los días que es su momento se presentaban complicados, pero sobre todo por la motivación que inducia en mi para poder finalizar esta etapa de mi vida. Le estoy muy agradecida.

A mis revisores, el Dr. Juan Gabriel Gonzales Serna, La Dra. Alicia Martínez Rebollar y Al Dr. Luis Gerardo Vela Valdés, quienes en conjunto desempeñaron una buena labor para que mi investigación mejorará con sus aportaciones.

A mi familia, la familia Miranda Castillo quienes estuvieron en todo momento motivándome para poder realizar mi trabajo. Pero en especial a mi hermano, Eduardo Castillo Rendón y mi padre, Raúl Castillo Miranda, quienes en ningún momento dejaron de creer en mí. Aunque en ocasiones pensará en desistir, siempre estuvieron a mi lado para poder seguir adelante.

A la familia Montes Martínez, quienes en cada momento estuvieron apoyándome y dándome ánimos para que pudiera culminar esta etapa.

A mi novio, Salvador Montes Martínez. Por acompañarme en este viaje. Por su apoyo incondicional no importando los estados de ánimo que yo presentará. Por su amor, su cariño, su confianza, su paciencia pero sobre todo por sus palabras de aliento que nunca faltaron cuando más lo necesitaba.

A la Dra. Leticia Sánchez, el Dr. Máximo López, la Lic. Karla Pichardo, el Lic. José Luis Carmona y el Lic. Fernando Martínez por su confianza y apoyo. Sus conversaciones y su ejemplo dejaron en mí, ganas de seguir luchando para alcanzar mis objetivos. Nuevamente al Dr. Juan G. Gonzales Serna y al Lic. Fernando Martínez por permitirme trabajar con ustedes en la estancia de residencia profesional. Dicha actividad fue la que me motivo a iniciar la maestría en esta institución.

A Alondra, Roberto Villarejo, Sadher, Felipe, German, Cristina, Abiud y Luis. Por su apoyo en las presentaciones y en los trabajos. Por sus palabras de aliento. Fueron grandes personas conmigo, muy buenos compañeros de tesis y sobre todo grandes amigos. Los quiero.

A mis amigos, Iris, Jatziri, Roxana, Bismark, Mateo, Yolitsi y Juan Carlos, por sus palabras de motivación y su confianza.

A los amigos que hice durante la maestría, Roberto Tamayo, Jorge Anaya, Jorge Lara, Rita, Alida, Sandro, Pedro, Domingo, Julia, Rodrigo, Hugo, Alejandra, Yolanda y Yahir, quienes han estado en los momentos de festejo pero sobre todo en los momentos difíciles. Les agradezco cada muestra de cariño. Muchas gracias por su amistad y su apoyo. Además de aquellos que no son alumnos pero que también me brindaron su amistad. A la Sra. Eli, la Sra. Sarita, a Roberto, a Gisela y a Lili, por mostrarme su cariño y por sus palabras de aliento.

# Resumen

La extracción de información semántica (extracción de hechos) se define como una subtarea de la extracción de información que se enfoca a la representación de datos estructurados en unidades llamadas hechos. La extracción de hechos en noticias puede ser utilizada para crear sistemas de noticias personalizadas e identificar sucesos de mayor interés. Además de contribuir a que una computadora adquiera conocimiento del tema tratado.

A pesar de las múltiples investigaciones acerca de la extracción de hechos, en el idioma español no existe un amplio estudio sobre este tema. En esta investigación se presenta un método basado en un análisis estadístico para identificar patrones lingüísticos asociados a hechos aplicados a noticias de desastres naturales escritos en español.

Se obtuvieron noticias de periódicos en línea de mayor circulación en México y se analizaron de manera morfosintáctica y estadística. La creación de una heurística permitió la identificación de patrones para la extracción de hechos.

Se tuvo como resultado un corpus con 1502 noticias de desastres naturales distribuidos en 8 categorías. La investigación presenta una precisión del 83% y una cobertura del 92% por lo que creemos que nuestro método de solución es aceptable para la extracción de información semántica.

La investigación está orientada a noticias de desastres naturales pero el método se puede aplicar en distintas categorías de noticias en las que se aplique un análisis estadístico.

# Abstract

The semantic information extraction (facts extraction) is defined as a subtask of extracting information that focuses on structured data representation in units called facts. The facts extraction in news can be used to create personalized news systems and identify occurrences of greater interest. In addition to contributing to a computer to acquire knowledge about specific subject.

In spite of the multiple investigations about facts extraction, there is not enough works about this topic. In this research, we present a method based on a statistical analysis to identify linguistic patterns associated with facts applied to natural disasters news written in Spanish.

The News were obtained from the most popular online newspapers in Mexico. As a result, we created a corpus with 1502 news of natural disasters, which is distributed in eight categories.

The news from the created corpus were analyzed statistically and in a morph syntactic way. Later based on the news analysis a heuristic was created which allows the pattern identification for the extraction of facts. The research presents an accuracy of 83% and 92% recall we believe that our method is accurate for the extraction of semantic information. Although this research is oriented to the natural disasters news this method can be applied to different news categories in which will be applied the same statistical analysis.

# Índice

# Índice de figuras

# Índice de tablas

# Índice de gráficas

# Capítulo 1
# Introducción

La información manejada por un humano es expresada mediante el lenguaje natural (denominación que se le da a la forma que el humano se expresa) y se encuentra escrita en revistas, periódicos, libros, informes, entre otros. Para que una computadora pueda entender esta información son necesarias técnicas que ayuden a la comprensión del lenguaje natural (Bolshakov & Gelbukh, 2000).

El procesamiento del lenguaje natural (PLN) es el campo que combina las tecnologías de la ciencia computacional (como la inteligencia artificial, el aprendizaje automático o la inferencia estadística) con la lingüística aplicada, con el objetivo de hacer posible la comprensión y el procesamiento asistidos por una computadora expresada en lenguaje humano para determinadas tareas, como la traducción automática, los sistemas de diálogo interactivos, el análisis de opiniones, entre otros.

Debido al gran manejo de datos, se han presentado complicaciones en la recuperación de información relevante. La extracción de información (IE *Information Extraction* acrónimo en inglés) forma un conjunto de textos expresados en lenguaje natural para extraer determinados conceptos de mayor interés. La extracción de hechos también conocido como **extracción de información semántica** se lleva a cabo mediante un conjunto de reglas para el estudio de patrones sintácticos, que ayudan a resolver diversas tareas relacionadas con la comprensión automática de texto.

En esta tesis se realizó una investigación acerca de la extracción de información semántica relacionada a temas sobre desastres naturales, ya que estos fenómenos ocurren con frecuencia en el país dando como resultado un número elevado de noticias.

## 1.1. Planteamiento del problema

El lenguaje de la noticia está condicionado por la finalidad de informar. Por ello, debe ser claro, breve, preciso, fluido, sencillo, ágil y fácilmente comprensible. La noticia narra el hecho sin dejar traslucir la opinión o actitud del periodista, ocurriendo en ocasiones que destaque el texto que más le interesa que el lector lea, siempre alterando la estructura establecida en la noticia (López Cubino, López Sobrino, & Bernabeu Morón, 2012).

A pesar de las múltiples investigaciones que se han realizado sobre el tratamiento de texto, relacionado a la extracción de información semántica (extracción de hechos), en el idioma español no se ha desarrollado de manera más amplia este tema, recayendo únicamente en libros de textos de nivel básico, textos técnicos, noticias financieras y en bibliografías de personajes históricos de México, pero no en noticias referentes a desastres naturales.

La investigación que se desarrolla en este trabajo, implementa un método para la extracción de información semántica contenida en noticias relacionadas con desastres naturales escritas en español. A través de este método se pueden llegar a beneficiar diversas tareas del Procesamiento de Lenguaje Natural, como la creación de resúmenes automáticos, el desarrollo de sistemas de pregunta-respuesta, la evaluación de la calidad de contenido de una noticia mediante el número de hechos encontrados con relación a su longitud, entre otros.

## 1.2. Justificación

La extracción de hechos permite a las técnicas del procesamiento de texto obtener información más comprensible para sistemas que requieren de información más precisa. Ayudando a sistemas de pregunta-respuesta, resúmenes automáticos o llenado de bases de conocimiento.

La extracción de hechos en noticias acerca de desastres naturales puede ser utilizada para crear sistemas de noticias personalizadas, proporcionando la selección de notas con mayor precisión conforme las preferencias de los usuarios y la identificación de sucesos de interés. Además contribuye a que una computadora adquiera conocimiento de desastres naturales.

Existen diversas investigaciones sobre la extracción de hechos principalmente en el idioma inglés, teniendo esto como desventaja en textos en español debido a que la estructura gramatical no es idéntica. Para el idioma español, las investigaciones con respecto a este tema son escasas, por lo cual esta investigación realiza una contribución al avance de la investigación del procesamiento de texto en español.

## 1.3. Objetivos

### 1.3.1. Objetivo general

Desarrollar un servicio web para extraer hechos de noticias en línea escritas en español sobre desastres naturales basado en reglas heurísticas y análisis estadístico, aplicado a un corpus de noticias y a los resultados de un analizador sintáctico.

### 1.3.2. Objetivos específicos

- Crear un corpus de noticias escritas en español sobre temas relacionados a desastres naturales.
- Realizar un análisis sintáctico de noticias extraídas.
- Crear un algoritmo y determinar heurísticas a utilizar para la extracción de hechos.
- Desarrollar un servicio web como base para pruebas.

# Capítulo 2
# Marco Teórico

El presente capítulo se explicará los conceptos esenciales y algunos conceptos secundarios que dan soporte al desarrollo de esta investigación.

## 2.1. Procesamiento del lenguaje natural (PLN)

El recurso más importante que poseen los seres humanos es el conocimiento. Durante toda la historia de la humanidad, el conocimiento se comunica, se guarda y se maneja en la forma de lenguaje natural (ya sea griego, latín, inglés, español u otros). En la época actual, el conocimiento sigue acumulándose en forma de documentos, libros, artículos, aunque ahora ya se guardan en forma electrónica, es decir digital.

Las computadoras son de gran ayuda para el procesamiento del conocimiento. Sin embargo, lo que es conocimiento para los seres humanos no lo es para las computadoras. Éstas copian, respaldan, transmiten o borran archivos, como un burócrata que pasa los papeles a otro sin leerlos. Pero no saben buscar respuestas a preguntas, hacer inferencias lógicas, generalizar y resumir un texto. Es decir, no hacen todo lo que las personas normalmente hacemos con el texto, porque no entienden (Gelbukh, 2010).

Para resolver dicha situación, en los países desarrollados se dedican esfuerzos al estudio del Procesamiento del Lenguaje Natural (PLN). Esta ciencia, en función de un enfoque práctico versus teórico, adopta varios nombres: procesamiento de lenguaje natural, procesamiento de texto, tecnologías de lenguaje o lingüística computacional. En todo caso, de lo que se trata es de procesar el texto por su sentido y no como un archivo binario, con el fin de lograr su comprensión.

## 2.2. Extracción de información

La extracción de información (IE por sus siglas en inglés) forma parte del procesamiento de lenguaje natural. Su propósito es obtener determinada información de un documento. Los sistemas de EI extraen eventos o hechos particulares de un dominio concreto y omiten aquellos que no lo son. Por ejemplo, una extracción en el dominio de noticias sobre fusiones de empresas debería obtener los nombres de las empresas que se fusionan, el capital de la inversión realizada, la actividad o producción de las empresas, su situación geográfica, entre otras (Català Roig & Castell Ariño, 1997).

## 2.3. Extracción de hechos

También denominada **extracción de información semántica**. Se considera una sub área de la extracción de información. Su función consiste en recuperar los hechos, enfocándose de manera más rígida a la representación de formas estructuradas.

### 2.3.1. Hechos (Extracción de Información Semántica)

Un hecho es la unidad mínima de una oración, que tiene dependencia semántica y contiene solo un verbo asociado a entidades y su forma es una tripleta conformada como:

**Hecho = [Sujeto] + [Verbo] + [Objeto/Complemento]**

Ejemplo:

"La Secretaría de Gobernación publicó este viernes la declaratoria de desastre natural para 21 municipios del estado de Guerrero que tuvieron afectaciones por el sismo de 6.4 grados en la escala de Richter" (ver Tabla 2.1).

**Tabla 2.1: Ejemplo de extracción de hechos en una noticia.**

| Sujeto | Verbo | Complemento |
|---|---|---|
| La secretaria de Gobernación | Publicó | La declaratoria de desastre natural para 21 municipios del Estado de Guerrero. |
| 21 municipios del Estado de Guerrero | Tuvieron | Afectaciones por el sismo de 6.4 grados en escala Richter. |

## 2.4. Niveles del lenguaje

El lenguaje está compuesto por cinco niveles (Fonológico, Morfológico, Sintáctico, Semántico y Pragmático), siendo cada uno de éstos fundamentales en la comunicación humana. Cada nivel tiene funciones específicas para el procesamiento lingüístico en determinados contextos.

A continuación, se describen solamente los niveles morfológico, sintáctico y semántico. Estos analizarán la información de los hechos durante esta investigación.

### 2.4.1. Morfología

La morfología, es la rama de la lingüística que estudia la estructura interna de las palabras para delimitar, definir y clasificar sus unidades, las clases de palabras a las que da lugar y formación de nuevas palabras (Hernando Cuadrado, 1995). Cumple tres funciones específicas: categoriza las palabras de acuerdo con su función (sustantivo, adjetivo, verbo, adverbio, entre otros); estudia las variaciones de sus formas; y explica los procesos que intervienen en la derivación y composición de las palabras.

La mayor parte de las palabras del español contienen una morfología formada por un lexema o raíz (**buen**-os); y unos gramemas o derivaciones (buen-**os**). El lexema indica el significado de la palabra y en ocasiones, su familia léxica: **café**: 'bebida que se hace con la semilla del cafeto'. Sus gramemas serían: **cafeto**, **cafetal**, **cafetería**, **cafetera**, **cafeína**.

Las diferentes modificaciones al significado de un lexema, se denominan prefijos (aparecen antes), sufijos (se colocan después) o infijos (se colocan en medio): **re**leer, jardin**ero,** herb**ol**ario.

### 2.4.2. Sintaxis

Etimológicamente sintaxis quiere decir "acción de disponer juntamente". Estudia la estructura de la lengua en cuanto a la combinación de las palabras para formar estructuras (Hernando Cuadrado, 1995). Es un conjunto de procedimientos o medios constructivos, como el orden de las palabras, el uso del artículo y de las preposiciones, de los tiempos verbales, pronombres, conjunciones, entre otros. Consiste en el estudio de las conexiones y significaciones gramaticales del decurso. Las unidades básicas de la sintaxis son el sintagma y la oración.

El **sintagma**, está formado a partir de la combinación de los monemas (lexemas y morfemas) en el decurso, es definido en la mayor parte de los tratos gramaticales como la unidad de función dentro de la estructura oracional, y se halla integrado por una función en abstracto (funlema) y por unos elementos concretos que la cubren (funtivos), sin que ello construya

obstáculo alguno, para que en determinados casos se encuentre con la presencia del signo 0 (Hernando Cuadrado, 1995).

Partiendo desde la definición de morfema, se puede definir el sintagma como una combinación de dos o más elementos significativos mínimos. Así en la oración "Los niños juegan" se tendría tres sintagmas, coincidiendo con el concepto de palabra. Pero en "Los niños juegan en su casa" se tendría cinco sintagmas. A saber: los, niños, juegan, su, casa.

Entre los enunciados existe un tipo especial conocido con el término de **oración** (Alarcos Llorach, 2000), también denominado como unidad básica de la sintaxis. Uno de sus componentes es el verbo (o sintagma verbal), contienen dos unidades significativas entre las cuales se establece la relación predicativa: el sujeto y el predicado, que se entienden tradicionalmente como "Aquello de que se dice algo" (sujeto), y "lo que se dice del sujeto" (predicado). Se comparan las siguientes oraciones, todas aplicadas a una misma situación y posibles repuestas de la pregunta "¿Qué hace el niño?":

El niño escribe en su cuarto una carta a su amigo.
El niño escribe una carta a su amigo.
El niño escribe una carta.
El niño escribe.
Escribe.

De una a otra oración se han eliminado datos, porque son conocidos por el interlocutor dado que son términos adyacentes cuya presencia no es indispensable para que pueda existir una oración.

Pero en todas ellas aparece la unidad escribe, imprescindible para que exista una oración. Esta forma verbal es el único de la oración, y en él se cumple la relación predicativa: se dice de alguien (la "tercera persona") algo (la noción de "escribir").

### 2.4.3. Semántica

El término semántica proviene del griego "semantikos", que significa "algo que tiene un significado relevante o significativo". La semántica es la parte de la lingüística que estudia el significado, la interpretación y el sentido de los signos lingüísticos, de las palabras, de los símbolos, de las expresiones y de sus combinaciones, sus formas gramaticales y sus cambios, así como su evolución en el tiempo. Este término fue creado por Michel Bréal en 1833.

Esta disciplina se ocupa del significado de los signos lingüísticos: palabras, oraciones y textos, "no estudia las unidades del nivel fónico, los fonemas y los sonidos, puesto que no tiene significado" (Mateos & Gonzalez, 2014)

## 2.5. Análisis morfosintáctico

El nivel morfosintáctico integra la morfología y la sintaxis. La forma en que se organizan y relacionan las palabras (sintaxis), así como la estructura de las mismas (morfología) aporta información útil e indispensable para la comprensión del significado de los enunciados.

En las secciones siguientes, se darán una explicación los procesos lingüísticos que se analizaran para el desarrollo de este trabajo.

## 2.6. Análisis morfológico

### 2.6.1. Lematización (stemming)

Consiste en determinar la forma, clase o categoría gramatical de cada palabra de una oración. Es una técnica de reducción, permite detectar variantes morfológicas de un mismo. Por ejemplo: palabras como computo, computadoras, computable, computación son variantes del termino computar, y reemplazadas por el termino raíz o lema (Tolosa, 2008).

El uso del stemming o lematización posibilita a:

a. Tener índices de menor tamaño.
b. Una mayor cantidad de respuestas a una consulta dada, debido a que ahora al aplicarse lematización al corpus y a la consulta se recuperan documentos que tengan todas las variantes morfológicas de los términos contenidos en la consulta.

Esta técnica permite extraer sufijos y prefijos comunes de tal forma que las palabras literalmente son diferentes pero tiene una raíz común. Se considera como un solo término en base a su raíz.

Por ejemplo: la palabra **perritas**, el lematizador determinará que se trata del lema “perro” con atributos femenino, plural y diminutivo. De manera similar si se usa la palabra **leerá**, indica la tercera persona del singular del futuro indicativo del verbo “leer”.

## 2.7. Análisis sintáctico

Su función es etiquetar cada uno de los componentes sintácticos que aparecen en la oración y analizar la combinación de palabras 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.

Las gramáticas, tal como se muestra a continuación, están formadas por un conjunto de reglas:

SN → sintagma nominal
SV → sintagma verbal
Det → determinante
O → SN, SV
SN → Det, N
SN → Nombre Propio
SV → V, SN
SV → V
SP → Preposición, SN

El resultado del análisis se puede mostrarse en forma arbórea. Los árboles son formas gráficas utilizadas para expresar la estructura de la oración, consistentes en nodos etiquetados (O, SN, SV) conectados por ramas.

### 2.7.1. Árbol de análisis sintáctico

Un sistema de lenguaje natural debe representar la estructura sintáctica de un modo que refleje cómo las palabras y sintagmas de una frase se relacionan entre sí. Dicha estructura sintáctica y la correspondiente relación entre palabras y sintagmas, se representan usualmente en una forma conocida como “árbol sintáctico”.

La estructura sintáctica actúa como una guía para la combinación de los pequeños trozos y frases que comprenden en una representación significativa de la frase completa. El árbol de estructura sintáctica puede proporcionar la información necesaria para la representación significativa puesto que incorpora conocimientos sobre qué palabras modifican a otras.

La generación de un árbol sintáctico es posible a la notación conocida como gramática de contexto libre. Es utilizada para representar el idioma inglés. Una gramática de contexto libre es una de los cuatro tipos de gramática definidos por Norman Chomsky en el MIT (Instituto Tecnológico de Massachusetts acrónimo en español) en la década de los cincuenta. Las gramáticas definidas por Chomsky van desde las muy generales a las muy restringidas con base en las reglas que permiten la sustitución de los componentes de la frase (por ejemplo, un

sintagma preposicional) por otros componentes sintagmáticos (tales como, una proposición enseguida de un sintagma nominal) en orden a generar frases correctamente estructuradas en un idioma.

### 2.7.2. Tokenización

Un *token* se puede definir como la unidad mínima de información con significado propio dentro de una secuencia de caracteres alfanuméricos. Estas cadenas de unidades mínimas de información o unidades léxicas, son generadas previamente por el módulo lexicográfico integrado en el analizador sintáctico encargado de identificarlas dentro de un texto o secuencia ordenada de caracteres alfanuméricos.

El proceso de *tokenización* consiste en la descomposición en forma de lista de esas cadenas de *tokens* en sus unidades mínimas. Así, un programa de este tipo podría generar la siguiente lista de *tokens* a partir de la frase "¡Hola Mundo!":

[161, 72, 111, 108, 97, 32, 77, 117, 110, 100, 111, 33]

Donde:

Cada uno de los números de la lista se corresponde con el carácter ASCII (*American Standard Code for Information Interchange* por sus siglas en inglés) correspondiente a cada una de las unidades mínimas de significación identificadas en la frase, en el mismo orden.

La *tokenización* es por tanto el proceso básico que permite manejar el lenguaje natural escrito para su posterior procesamiento, en base a su descomposición en unidades mínimas de información con significado propio.

### 2.7.3. Analizador sintáctico

Es la fase del análisis que se encarga de checar la secuencia de *tokens* que representa al texto de entrada, con base en una gramática dada. En caso de que el programa de entrada sea válido, suministra el árbol sintáctico que lo reconoce a partir de una representación computacional. Este árbol es el punto de partida de la fase posterior de la etapa de análisis: el analizador semántico (Gálvez Rojas & Mora Mata, 2005).

El analizador sintáctico dirige el proceso de compilación, de manera que el resto de las fases evolucionan a medida que el analizador va reconociendo la secuencia de entrada de modo que, a menudo el árbol ni siquiera se genera realmente.

El analizador sintáctico también:

- Incorpora acciones semánticas en las que colocar el resto de fases del compilador (excepto el analizador léxico), desde el análisis semántico hasta la generación de código.
- Informa de la naturaleza de los errores sintácticos que encuentra e intenta recuperarse de ellos para continuar la compilación.
- Controla el flujo de *tokens* reconocidos por parte del analizador léxico.
- Realiza casi todas las operaciones de la compilación, dando lugar a un método de trabajo denominado compilación dirigida por sintaxis.

## 2.8. Heurística

La palabra heurística procede etimológicamente de la palabra griega “euriskein” que procede de “eureka”, un vocablo que significa hallar o encontrar. El diccionario de la Real Academia Española (Real Academia Española) define a la palabra heurística como:

- Técnica de la indagación y el descubrimiento.
- Búsqueda o investigación de documentos o fuentes históricas.
- En algunas ciencias, manera de buscar la solución de un problema mediante métodos no rigurosos, como por tanteo, reglas empíricas, etc.

## 2.9. Noticia

La noticia es el relato oral o escrito de un suceso interesante y actual. Es la información de un hecho de interés ocurrido recientemente. Constituye el elemento primordial de la información periodística y el género básico del periodismo.

Un acontecimiento se convierte en noticia por alguna de estas razones: por su actualidad, su proximidad al lector, su trascendencia, su relevancia, su capacidad para emocionar, los conflictos que plantea, su rareza, etc.

La noticia se caracteriza por la objetividad, la claridad y la concisión. En ella no aparecen nunca comentarios u opiniones. Su contenido responde, siempre que es posible, a estas seis preguntas clásicas: ¿quién?, ¿qué?, ¿cómo?, ¿cuándo?, ¿dónde?, ¿por qué?

La noticia presenta una estructura precisa, formada por el titular y el cuerpo. El **titular** es el elemento más relevante de una noticia. Se compone de uno o varios elementos: cintillo (clase especial de título a toda plana y de una sola línea que encabeza una página), antetítulo (enmarca la noticia), título o cabeza (introduce o presenta la noticia), subtítulo (añade alguna particularidad) y sumario (indicación o resumen breve del contenido de la noticia). El **cuerpo** desarrolla el contenido de la noticia presentando los hechos y la información de otros elementos complementarios que ayudan a una mejor comprensión del acontecimiento (López Cubino, López Sobrino, & Bernabeu Morón, 2009).

# Capítulo 3
# Estado del Arte

En esta sección se presentan algunos trabajos relacionados con la extracción de hechos que influyeron para documentar esta investigación elaborando una tabla comparativa al término de este apartado, permitiendo analizar y evaluar los trabajos que fundamentan este documento.

## 3.1. Sistema de extracción de hechos de las personalidades históricas de México

La investigación desarrollada por Hernández Jiménez, Marines Zane, & Montes San Agustín (2012), muestra la extracción de hechos sobre personajes históricos en México.

Las investigaciones en relación a las técnicas como el procesamiento del lenguaje natural y la extracción de hechos, no han sido suficientes para el diseño de un sistema que interprete el lenguaje de un humano. La existencia de grandes volúmenes de datos en formato electrónico referentes a personajes históricos de México y la falta de análisis por su insuficiencia estructuración, dieron inicio a esta investigación.

Por tal motivo, plantean solucionar el problema con base en plantillas (Fichas de información) de extracción de hechos bajo una herramienta precisa que muestre a los alumnos de nivel básico, información clara y una mejor compresión de cada personaje. Se implementan reglas heurísticas para la extracción de los patrones con el objetivo de obtener un rango elevado de precisión y ser de utilidad en el nivel educativo básico.

El procedimiento para la solución del problema se conforma por las siguientes fases: Etiquetado, Pre procesamiento, Descubrimiento de Patrones de extracción y Generación de Plantillas.

La fase de **preprocesamiento** realiza la *tokenización*, el análisis morfosintáctico y el reconocimiento de entidades nombradas. El **descubrimiento de patrones de extracción** lleva a cabo un proceso de correspondencia para el desarrollo de una plantilla. Por último **la generación de plantillas** indica el llenado de un modelo determinado con la información extraída de los datos presentados de cada personaje.

Como resultado presenta la plantilla con las reglas establecidas más los hechos obtenidos mostrando un 80% de precisión.

## 3.2. Algoritmo heurístico para la extracción de hechos de uso modelo relacional y datos sintácticos

Esta investigación de Sidorov & Herrera-de-la-cruz (2011) presenta el desarrollo de un algoritmo heurístico para realizar extracción de hechos aplicados a textos estructurados.

El algoritmo creado extrae los hechos de oraciones usando una representación basada en un modelo de datos relacional. Estas oraciones son seleccionadas de libros de textos estructurados empleando un analizador sintáctico llamado *Connexor*. La información analizada sirve para la resolución de correferencia y de reconocimiento de entidades nombradas. Los resultados son presentados en un documento XML (*eXtensible Markup Language* por su acrónimo en inglés) que contiene elementos textuales (por ejemplo: párrafo, oración, palabra, *token*).

La investigación muestra una desventaja hacia el analizador sintáctico *Connexor* debido a que es más óptimo para analizar texto en inglés. Los autores sugieren utilizar *Freeling* para realizar análisis sobre texto en español.

Sidorov & Herrera-de-la-cruz (2011) mencionan un inconveniente al utilizar lógica de primer orden como representación formal ya que solo serán utilizados por aplicaciones que entiendan este tipo de técnicas. Los resultados expresan un total del 80 por ciento de precisión en las oraciones procesadas.

## 3.3. Extracción automática de hechos en libros de textos basada en estructuras sintácticas

La investigación presentada por Aguilar Galicia, Sidorov, & Nikolaevna Ledeneva (2012) señalan una extracción de hechos aplicados a textos estructurados educativos de nivel primaria y secundaria en español.

La problemática plateada por los autores indica que no existe sistema para la extracción de hechos desde textos en español, eligiendo libros de educación básica porque tiene áreas de estudio bien definidas y contienen una gran cantidad de hechos.

Para solucionar el problema, emplearon un método de solución obteniendo la siguiente arquitectura: libros de texto (libros de nivel primaria y secundaria), pre procesamiento, análisis sintáctico, extracción de hechos y almacenamiento de hechos extraídos.

En la fase de libros de texto, se eligen lecciones y se selecciona un conjunto de oraciones. Bajo un analizador sintáctico se realiza un análisis morfológico y sintáctico para obtener un árbol de dependencias. Mediante un conjunto de heurísticas se analizan los arboles de dependencias y se extraen los hechos de cada oración por consecuente se almacenan en una base de datos.

Como resultado presentan un corpus integrado por 68 oraciones, extrayendo hechos en solo 59 oraciones ya que el analizador no permitió etiquetar correctamente las oraciones. Obtuvieron un total de 157 hechos, con 137 de ellos correctos manteniendo un 87 por ciento en la precisión.

## 3.4. Extracción automática de hechos de los comunicados de prensa para generar noticias históricas

El trabajo de (Andersen et al., 1992) describe un sistema de tipo JASPER (Asistente Periodista para la Elaboración de Informes de Resultados), desarrollado por el Grupo Carnegie para Reuters Ltd.

Los autores indican que la aplicación JASPER está basado en plantillas y técnicas de compresión para la extracción de hechos de un fragmento de texto seleccionado. JASPER toma comunicados de prensa de transmisión en vivo de empresas. Identifica la información sobre las utilidades, los dividendos y los lanzamientos y extrae un conjunto predeterminado de información.

La investigación muestra que el sistema realiza una combinación de representaciones basadas en marcos de conocimiento, procedimiento orientado a objetos, patrones y heurísticas

con la finalidad de aprovechar las convenciones estilísticas como regularidades léxicas, sintáctica, semánticas y paradigmas que son observados en el corpus del texto. Logrando buenos resultados de precisión. Cabe señalar que Reuters proporciona noticias financieras en tiempo real a operadores financieros.

Estadísticas emitidas del autor sobre su investigación indica que JASPER maneja el 33 por ciento de las comunicaciones a la perfección, el 21 por ciento de las historias sobre ganancias han marcado sin errores y omisiones.

La investigación indica que el sistema puede aplicarse provechosamente a aplicaciones comerciales reales. La información debe ser seleccionada cuidadosamente para obtener resultados positivos.

## 3.5. Extracción automática de los hechos, relaciones y entidades para la web a gran escala de la población base de conocimientos

Cuando la fuente de texto es la Web los métodos de extracción deben hacer frente a la ambigüedad, el ruido, la escala y actualizaciones. El objetivo de la tesis de Nakashole (2013) fue desarrollar métodos que permitieran la creación automática de bases de conocimiento de características mencionadas de textos Web.

La tesis hace tres contribuciones: La primera contribución es un método para la minería de hechos a escala de alta calidad, a través del razonamiento de restricción distribuido y un modelo de representación del patrón que es fuerte frente a patrones ruidosos. La segunda contribución es un método para la extracción de una gran colección completa de tipos de relación más allá de los que comúnmente se encuentran en las bases de conocimiento existentes. La tercera contribución es un método para la extracción de los hechos a partir de fuentes web dinámicas, tales como: artículos de noticias y medios de comunicación social, donde uno de los principales retos es la aparición constante de nuevas entidades.

El proceso de extracción de hechos, toma como entrada un corpus de documentos de texto de lenguaje natural estructurado y produce datos estructurados en forma de la tripleta “sujeto - verbo – predicado”. Se presenta un sistema denominado PROSPERA mediante el cual, utiliza la técnica de n-gramas además emplea algoritmos para el cálculo de la similitud en patrones y generar los hechos candidatos. El autor muestra un resultado del 83% de precisión.

## 3.6. Extracción automática de hechos clave de documentos individuales en artículos de prensa

El trabajo de Kastner & Monz (2009) aborda la extracción de los hechos de artículos de prensa. Utiliza características estadísticas, sintácticas, semánticas y generales para identificar las frases más importantes de un documento.

En este trabajo se enfoca en identificar las oraciones relevantes de la noticia. El sistema implementado es llamado AURUM *(AUtomatic Retrieval of Unique information with Machine learning)* orientado en características como: longitud de la oración, la frecuencia n-gramas, posición de sentencia, la identificación adecuada del sustantivo, similitud con el título, tf-idf.

Su *corpus* está compuesto de 1200 artículos de las noticias CNN, seleccionados de una amplia gama de temas (política, negocios, deportes, salud, temas del mundo, clima, entretenimiento y tecnología) aunque solo se consideraron los artículos con aspectos más destacados en la nota, con el fin de determinar las frases que se destacan más en la historia.

El método que toma para la extracción de hechos clave es un modelo probabilístico que permite determinar la posición y apariciones de las frases que destaquen en la oración. Kastner & Monz (2009) presenta el 80% de precisión como resultados de su investigación.

## 3.7. Extracción de hechos para textos en el idioma polaco

El artículo de Boiński & Brzeski (2014) demuestra la extracción de hechos de textos en el idioma polaco con el uso de la enciclopedia web *Wikipedia*. A comparación del inglés, el idioma polaco tiene más plantillas de conjugación por lo que complica más la inflexión.

El método presentado en la investigación de Boiński & Brzeski (2014) extrae hechos en forma de triplete (sujeto - predicado - objeto) con la selección de textos con Pantera y extrae entidades nombrada con *Nerf.* Los verbos u entidades nombradas de tipo: nombres personales, topónimos, nombre de organizaciones y nombres geográficos son asignados a sujetos y objetos. Entidades de tipo fecha se asignan a objetos, los verbos en participio y preposiciones se asignan al predicado.

La asignación de los verbos y entidades nombradas a sujetos y objetos en tripletas depende del contexto morfosintáctico que es utilizado. La investigación presenta los siguientes problemas:

El etiquetador morfosintáctico requiere una base de datos de entidades nombradas y sus formas básicas para formular correctamente dada entidad.

Existen inconvenientes al detectar las siglas y verbos causado por la dificultad de adivinar la forma correcta del lema.

(Boiński & Brzeski, 2014) indican que los inconvenientes pueden resolverse con *Słowosieć* (una versión polaca de WordNet).

## 3.8. DEBORA: Extracción de tripletes entidad-relación basado en dependencias de textos polacos de dominio abierto

El análisis de texto polacos habían sido un problema ya que se caracteriza por su alto grado de marcado morfológico y flexibilidad en el orden de palabras. La investigación de Wróblewska & Sydow (2012) se centra en la extracción automática de tripletas entidad-relación a partir de un corpus generado por textos web. El objetivo de la investigación es realizar la extracción utilizando técnicas basadas en la dependencia.

El método de extracción se aplica a un conjunto de 188,415 artículos de noticias web polacas, dividiéndose en 6, 303,794 frases con un promedio de 20 tokens por oración, analizando 3, 265,817 oraciones que contaban con una entidad nombrada o con una entidad reconocida.

Wróblewska & Sydow (2012) realizan dos evaluaciones para llevar a cabo el proceso de extracción. La primera evaluación, calcula la aproximación de precisión mediante el muestreo de 100 tripletes al azar y de forma manual se examina su valides. Se repitió el cálculo tres veces hasta que se logró una precisión media de 54%. Seleccionando los tripletes que representan relaciones relativas a las personas y lugares como nacimiento, fallecimiento y el lugar donde vive.

Para la segunda evaluación, se prepara un texto para probar el algoritmo DEBORA, que consta de 64 frases sencillas con un promedio de 9 tokens por oración, teniendo como resultado una presión de 95.8% y cobertura de 92%.

## 3.9. Modelos estructurales, transitivos y latentes de extracción de hechos biográficos

Este artículo muestra seis enfoques para realizar la extracción de hechos bibliográficos de modelos estructurales, transitivos y latentes. Proponiendo una mejora del trabajo realizado por Deepak Ravichandran y Eduard Hovy.

(Garera & Yarowsky, 2009) proponen y evalúan los siguientes enfoques:

1. Realizar una versión mejorada de la investigación realizada por Deepak Ravichandran y Eduard Hovy "Algoritmo basado en modelos *Untethered* de diagrama parcialmente contextuales".
2. El aprendizaje de un modelo basado en posiciones absolutas, relativas y el orden de hipótesis secuencial que satisface el modelo de dominio, dando como ejemplo la fecha de fallecimiento por encontrase enseguida de la fecha de nacimiento en una bibliografía.
3. Uso de modelos transitivos sobre los atributos a través de las entidades concurrentes.
4. Utilizar modelos para detectar atributos que no se pueden mencionar en un artículo.
5. Uso de correlaciones interatributarias para filtrar combinaciones de atributos biográficos improbables.
6. Aprender las distribuciones de las funciones de atributos, por ejemplo, el uso de un distribución edad filtrar tripletes que contienen improbables <año de fallecimiento> <año de nacimiento> valores de vida útil.

Para la implementación se empleó un patrón modelo estándar de aprendizaje. Los artículos fueron extraídos de Wikipedia. Nombre de las persona, no se puedo obtener de Wikipedia debido a la información incompletas y no normalizadas, por lo que utilizaron ejemplos de una base de datos bibliográfica en línea llamado NNDB (*Notable Names Database*, acrónimo en inglés), extrayendo los siguientes atributos: "fecha de nacimiento", "lugar de nacimiento", "fecha de muerte", "sexo", "nacionalidad", "ocupación" y "religión".

Este trabajo presento una mejora a la norma realizada por Ravichandran y Hovy utilizando patrones contextuales *Untethered*, seguidos por un enfoque de una posición de documento y la secuencia basada en atribuciones de modelado. Se logró el 80% de precisión media en una prueba del conjunto de atributos bibliográficos.

## 3.10. Extracción de tripleta de oraciones usando SVM

La investigación de Dali & Fortuna (2008) presenta la extracción de tripletas de oraciones escritas en inglés, usando el método SVM (*Support Vector Machine* acrónimo en inglés).

Dali & Fortuna (2008)que los datos se originan de notas realizadas por lingüistas en un corpus de noticias publicadas por Reuters en el año 2000. El procedimiento que se sigue para identificar los tripletes es tokenizar la oración y eliminar stopwords. Las palabras sobrantes se realizaran combinaciones posibles ordenando de tres palabras en una lista, estos se los denominaron **candidatos a tripletes**. Posteriormente se hará uso al modelo SVM para asignar puntuación positiva a aquellos candidatos que serán extraídos y una puntuación negativa a los sobrantes. Las pruebas se realizan con 700 oraciones, contando con una precisión de 38.36% y cobertura 46.80%

## 3.11. Tabla comparativa de trabajos relacionados

En la siguiente tabla se mostrará una comparativa de los trabajos relacionados que se estudiaron en esta tesis. Las columnas corresponden a características que se tomaron como puntos de comparación.

**Tabla 3.1: Tabla de trabajos relacionados.**

| Trabajos relacionados | Fuente de extracción de Hechos | Idioma | Precisión | Recursos Utilizados | Descripción |
|---|---|---|---|---|---|
| Sistema de Extracción de Hechos de las personalidades Históricas de México. | Personajes Históricos de México. | Español | 80% | No especificado | El sistema extrae hechos sobre información de personajes históricos de México, obteniendo información relevante y estructurada con respecto a documentos bibliográficos de dichos personajes. |
| Algoritmo heurístico para la Extracción de Hechos de Uso modelo relacional y sintáctico de datos. | Textos técnicos o educativos. | Español | 80% | Analizador Sintáctico (Connexor) | Los algoritmos realizados en esta investigación se realizaron con fines de extraer hechos mediante textos técnicos y educativos |
| Extracción automática de hechos en libros de textos basada en estructuras sintácticas. | Oraciones seleccionadas de libros de texto de educación primaria y secundaria. | Español | 87% (Cobertura 90%) | Analizador Sintáctico (FreeLing-2.2) | Mediante oraciones de textos de libro de nivel básico seleccionadas por el usuario, se extraen los hechos y se almacenan en una base de datos. |
| Extracción Automática de Hechos de los Comunicados de Prensa | Información sobre utilidades, dividendos, | Inglés | 96% | No especificado | Está basada en plantillas y técnicas de comprensión. La aplicación toma comunicados de prensa de transmisión en vivo de empresas, |

| | | | | | |
|---|---|---|---|---|---|
| para generar Noticias Históricas. | lanzamientos de una empresa. | | | | identifica información sobre utilidades, dividendos, lanzamientos y extrae los hechos. |
| Extracción automática de los hechos, relaciones y entidades para la web a gran escala de la población base de conocimientos. | Extracción de hechos en Texto Web. | Ingles | 83% | No especificado | Se realiza una investigación acerca de la extracción de los hechos a partir de fuentes de web dinámicas, tales como: artículos de noticias y medios de comunicación social. |
| Extracción automática de hechos clave de documentos individuales en artículos de prensa. | Extracción de hechos en noticias de la prensa CNN. | Ingles | 80% | No especificado | Se enfoca a la longitud de la oración, la frecuencia n-gramas, posición de sentencia, la identificación adecuada del sustantivo, similitud con el título, tf-df, aplicadas a las noticias del periódico CNN escrito en inglés. |
| Extracción de hechos para textos en el idioma polaco. | Textos polacos de la enciclopedia web Wikipedia. | Polaco | No especificado | Lematizador (Pantera) Extracción de nombres de entidades (Nerf) | Se emplea la extracción de hechos en documentos de Wikipedia en el idioma polaco. |
| DEBORA: Extracción de triples entidad-relación basado en dependencias de textos polacos de dominio abierto. | Textos web polacos. | Polaco | 95.8% Cobertura 92% | No especificado | Esta investigación se centra en la extracción de tripletas entidad relación basado en técnicas de dependencia de los textos polacos. |
| Modelos estructurales, transitivos y latentes de extracción de hechos biográficos. | Textos de Wikipedia | Ingles | 80% | No especificado | Centra su investigación en 6 enfoques a la extracción de hechos biográfico utilizando estructural, transitiva y propiedades latentes de datos biográficos, realizando una |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | mejora a la norma Ravichandran y Hovy. |
| Extracción de tripleta de oraciones usando SVM. | Corpus de noticias publicadas por Reuters. | Ingles | 38.36% Cobertura 46.80% | No especificado | Extracción de tripletas de oraciones escritas en inglés, usando el método SVM. |
| Extracción Automática de Hechos de Noticias de Desastres naturales escritas en español (Investigación Propuesta). | Noticias escritas referidas a desastres naturales escritas en español. | Español | **Precisión 82% Cobertura 92%** | Analizador Sintáctico (Freeling) | Extracción de hechos de noticias de desastres naturales escritas en español, basado en reglas heurísticas y análisis estadísticos aplicados a un corpus de noticias y a los resultados de un analizador sintáctico. |

# Capítulo 4
# Método de solución

Este capítulo presenta el método de solución propuesto en esta investigación. El punto 4.1 muestra un diagrama y su descripción de las fases que lo componen. Además indica los algoritmos desarrollados basándose en los patrones sintácticos y ejemplos de hechos extraídos con estos patrones.

## 4.1. Descripción general

Para alcanzar el objetivo de la extracción de hechos, se desarrolló un método de solución. A continuación se muestra un diagrama y se explican las fases que lo conforman.

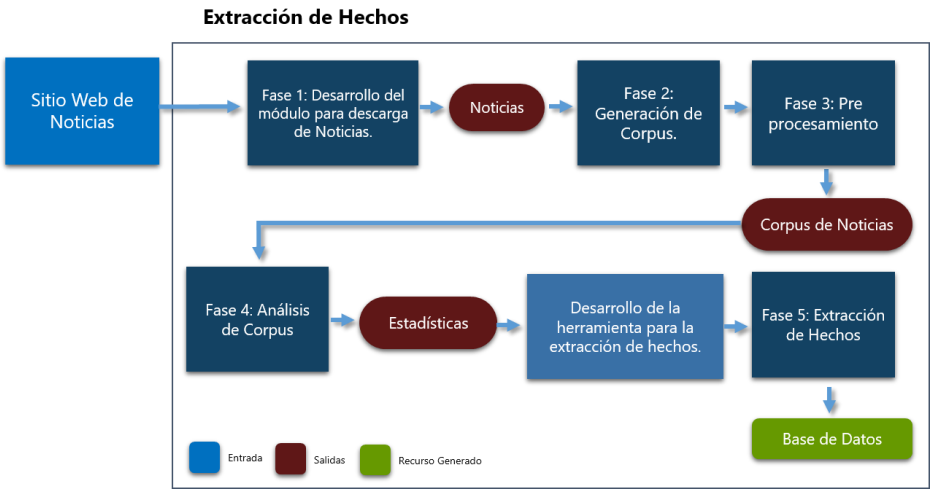



**Figura 4.1: Método de solución.**

**Fase 1: Desarrollo del módulo de descarga de noticias**

Esta actividad consiste en el desarrollo de un módulo que permita la descarga de noticias. El módulo se realizó en el lenguaje de programación *Java*. Este implementa una técnica denominada *Scraping* (técnica para la obtención de texto que se encuentra en páginas HTML) con la ayuda de la librería *Jsoup* (librería que se utiliza en aplicaciones Java para la obtención de texto de páginas web) para obtener las noticias de los periódicos en línea seleccionados. El conjunto de noticias descargadas se almacenan en archivos de texto plano.

**Fase 2. Generación del corpus**

Las noticias que se descargan en la fase 1 componen el corpus. Éste se conforma por 8 categorías (ciclones, heladas, precipitación pluvial, sismos, sequía, tormentas, inundaciones y tsunami) determinadas por el Centro Nacional de Prevención de Desastres (CENAPRED) (Centro Nacional de Prevención de Desastres, 2001) y Hofliger, Mahul, Ghesquiere, & Perez, 2012, formando un corpus con un total de 1502 noticias a procesar.

**Fase 3. Preprocesamiento**

Esta fase se enfocó al análisis morfosintáctico de la noticia. Se unen cada una de ellas, colocando un delimitador al inicio y otro al final para identificar una noticia con otra. Posteriormente se lematizan las palabras con el analizador sintáctico *Freeling* (Apéndice A).

**Fase 4. Análisis del corpus**

Se implementó un módulo que genera la frecuencia de verbos, palabras, gramas, número de noticias analizadas, promedio de oraciones por noticia y el promedio de palabras por oración. Estos resultados son almacenados en archivos de texto generados por cada categoría. Así mismo, se genera un archivo de texto con el conteo del número de las noticias analizadas, el promedio de oraciones existentes por categoría y el promedio de palabras que se encuentran por oración de cada categoría.

**Fase 5. Extracción de hechos**

Para iniciar esta fase se extraen patrones con la finalidad de comprobar la eficiencia de un método (análisis estadístico) diferente. Se generó una heurística para extraer patrones que toma ciertas palabras denominadas palabra base o pivote (son aquellas palabras en las que es posible encontrar hechos en su contexto por ejemplo términos frecuentes, verbos frecuentes, palabras asociadas, entre otros).

La heurística se dividió en cuatro variantes según el tipo de pivote que toma. Esto es, la variante 1 toma como palabras base los verbos frecuentes. La variante 2, utiliza como palabras base los verbos defectivos asociados a las categorías. En la variante número 3, sus palabras base

son los términos frecuentes y en la variante 4, se identifica el patrón con el uso de palabras asociadas es decir, las palabras afín de cada categoría**.**

## 4.2. Fase 1: Desarrollo del módulo para descarga de noticias

El proceso de esta actividad consiste en el desarrollo de un módulo que realice la descarga de noticias considerando los periódicos populares en circulación.

El módulo se realizó en el lenguaje de programación *Java*. Éste realiza una técnica denominada *Scraping* con la ayuda de la librería *Jsoup* (Jonathan Hedley, n.d.). En este caso, se usa para adquirir texto publicado de los sitios web de periódicos seleccionados en línea.

Para aplicar la librería *Jsoup*, se utiliza un archivo de configuración en formato XML que contiene los siguientes elementos:

- **<nombre>** (Núm. 1 de la Figura 4.2): Su contenido es el nombre del periódico seleccionado. Por ejemplo: El Universal, La jornada, CNN México, entre otros.
- **<sección>** (Núm. 2 de la Figura 4.2): Su contenido determina la clasificación que pertenece la noticia como: ciclones, heladas, precipitación pluvial, tormenta, inundaciones, sismos, sequía y tsunamis.
- **<url>** (Núm. 3 de la Figura 4.2): Contiene el enlace de cada noticia.
- **<contenedor>** (Núm. 4 de la Figura 4.2): Esta etiqueta contiene una notación CSS que indica la sección de la lista donde se encuentra la URL de cada noticia.
- **<css_titulo>** (Núm. 5 de la Figura 4.2): Contiene la clase CSS donde se encuentra el título.
- **<css_contenido>** (Núm. 6 de la Figura 4.2): Su contenido es el id o clase CSS del que se extrae el contenido de la noticia.

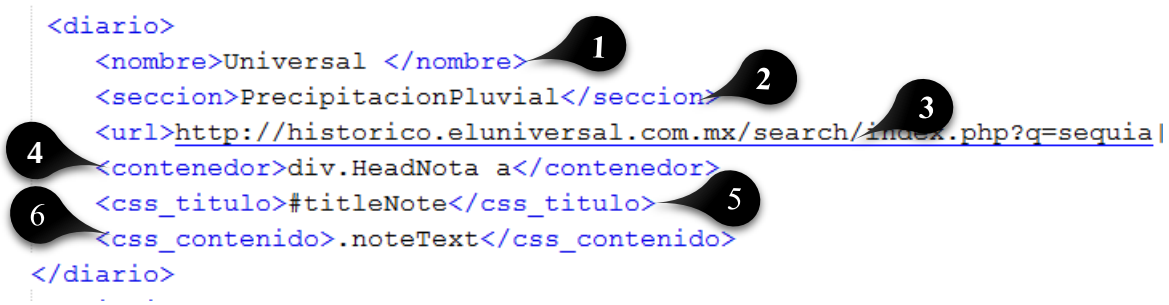



**Figura 4.2: Archivo xml de configuración para scraping.**

Posteriormente configurado el archivo XML se ejecuta la aplicación. Para iniciar la descarga de noticias se da clic en el botón **iniciar descarga** (Figura 4.3). La aplicación, muestra la descarga de noticias indicando los sitios web donde se obtienen.

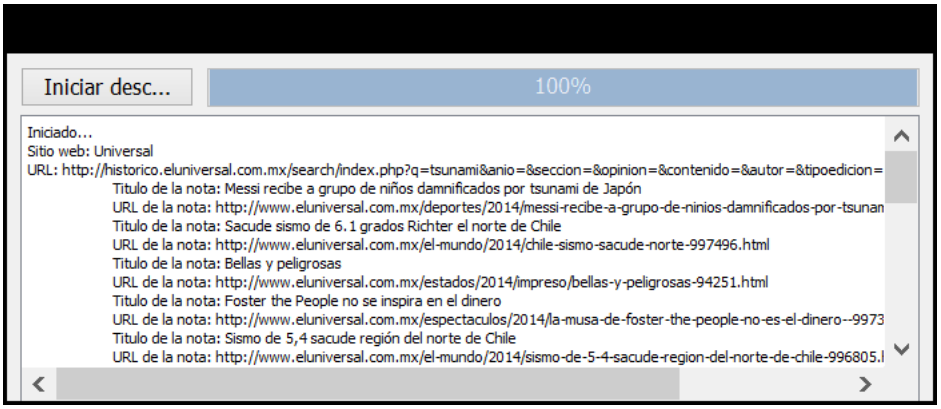



**Figura 4.3: Vista de la aplicación para la obtención de noticias**

El conjunto de noticias descargadas se almacenan en archivos de texto plano en la categoría que se le indica en el archivo XML (Núm. 2 de la Figura 4.2).

## 4.3. Fase 2: Generación de corpus

Las noticias que se descargan en la fase 1 conforman el corpus de entrenamiento. La aplicación almacena las noticias en una carpeta denominada “Noticias” en las categorías correspondientes en archivos de texto plano así como se muestra en la figura.

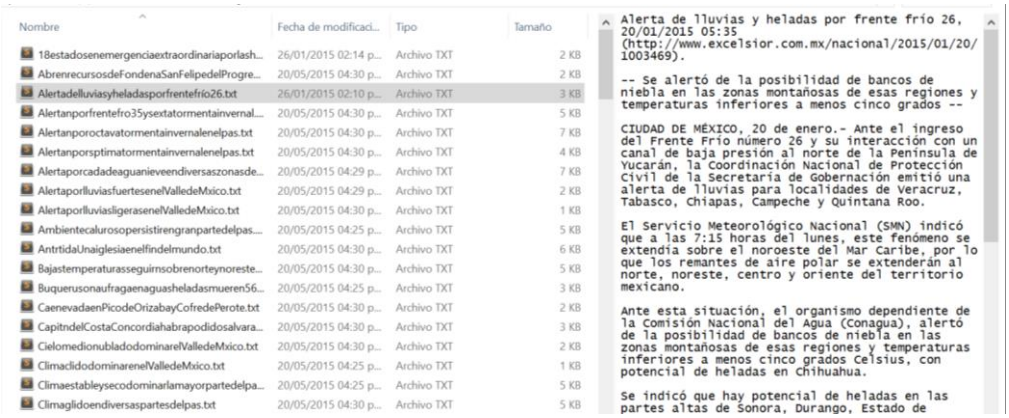



**Figura 4.4: Lista de noticias de la categoría heladas que forman el corpus.**

El archivo de texto plano de cada noticia contiene de los siguientes elementos:

- **Título de la noticia (Núm. 1 de la Figura 4.5):** Este texto se encuentra en la primera línea de cada archivo.
- **Enlace de la noticia (Núm. 2 de la Figura 4.5):** Seguido del título, se encuentra el enlace donde se obtiene la noticia.
- **Cuerpo de la noticia (Núm. 3 de la Figura 4.5):** Después de los dos elementos anteriores, se encuentra el contenido de la noticia. En éste sólo se encuentra el contenido escrito de la noticia, no tiene otros elemento que no se hayan configurado en el archivo XML.

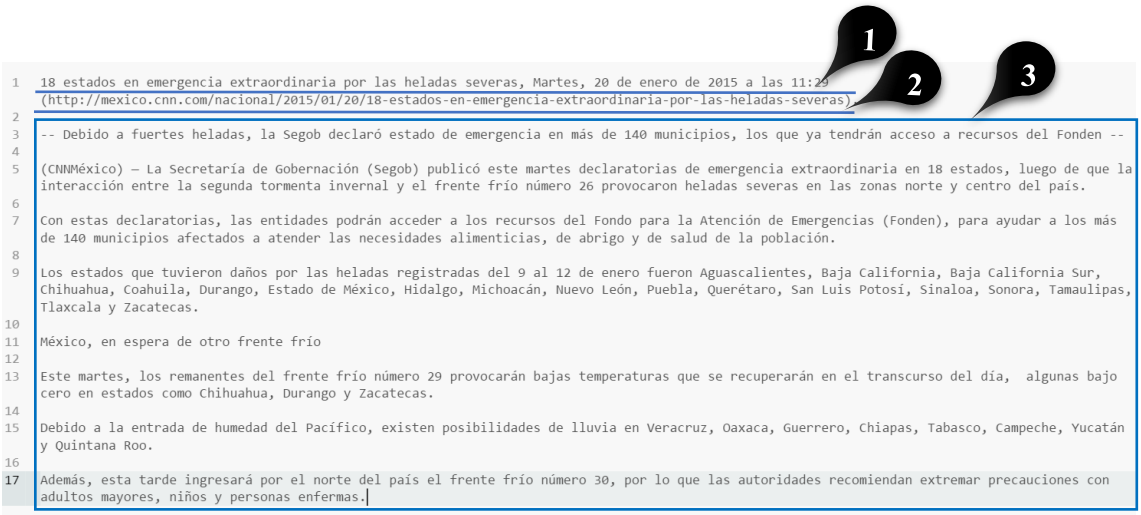



**Figura 4.5: Contenido del documento.**

El corpus de noticias está formado por 8 categorías determinadas por el Centro Nacional de Prevención de Desastres (CENAPRED) (Centro Nacional de Prevención de Desastres, 2001) y Hofliger et al., 2012, ya que son los fenómenos naturales que se informan con frecuencia en los periódicos. Estas categorías son:

- Ciclones
- Heladas
- Precipitación Pluvial
- Sismos
- Sequía
- Tormentas
- Inundaciones
- Tsunami

La Tabla 4.1: Tabla de categorías con el total de noticias encontradas en el corpus de entrenamiento. muestra el total de noticias por cada categoría del corpus de entrenamiento.

**Tabla 4.1: Tabla de categorías con el total de noticias encontradas en el corpus de entrenamiento.**

| Categorías | Total de noticias en el corpus |
|---|---|
| Ciclones | 215 |
| Heladas | 121 |
| Precipitación pluvial | 162 |
| Sismos | 149 |
| Sequía | 141 |
| Tormentas | 270 |
| Inundaciones | 230 |
| Tsunamis | 214 |

## 4.4. Fase 3: Preprocesamiento

En esta fase se lleva a cabo un análisis morfosintáctico de la noticia. El corpus de noticias está compuesto por 1502 noticias, es decir, 1502 archivos de texto. Para un óptimo análisis de las noticias, se unen etiquetando cada una de ellas, colocando un delimitador al inicio (núm. 1 de la Figura 4.6 ) y otro al final de cada noticia (núm. 2 de la Figura 4.6).

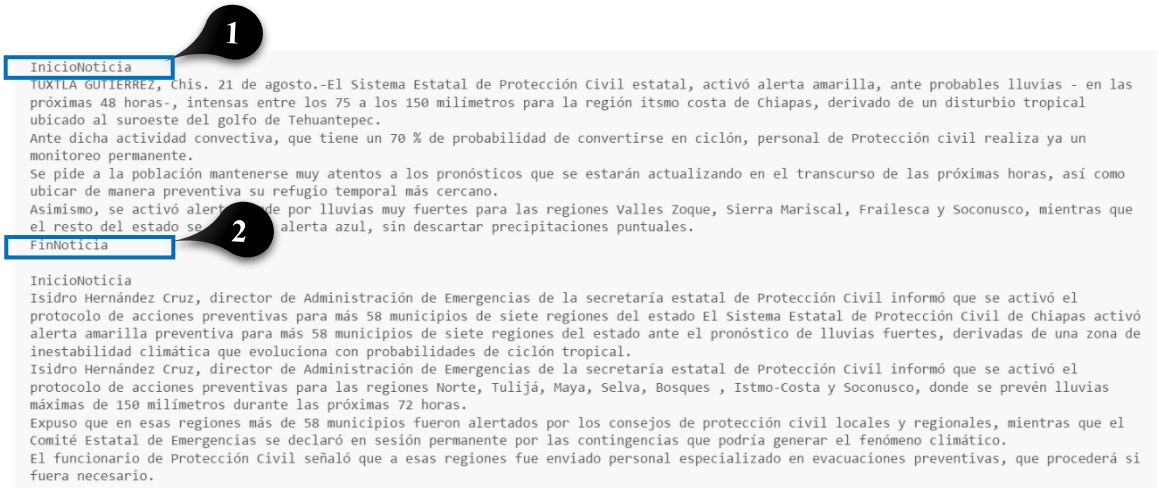



**Figura 4.6: Etiquetado de noticias.**

Posteriormente se lematizan las palabras con el analizador sintáctico *Freeling* (ver Figura 4.7) que otorga la forma de la palabra, su lema, la etiqueta que indica la información morfológica de las palabras a la que pertenece y el puntaje de aproximación del lema a la forma de palabra para que su análisis sea fácil al aplicar técnicas para la extracción de hechos.

```
InicioNoticia inicionoticia NP00000 1

-- -- Fg 1
Los el DA0MP0 0.976481
Cabos cabos NP00000 1
sufrió sufrir VMIS3S0 1
una uno DI0FS0 0.951575
de de SPS00 0.999984
sus su DP3CP0 0.999692
peores peor AQ0CP0 1
temporadas temporada NCFP000 1
ciclónicas ciclónico AQ0FP0 1
, , Fc 1
a a SPS00 1
el el DA0MS0 1
pronosticar pronosticar VMN0000 1
CONAGUA conagua NP00000 1
un uno DI0MS0 0.987295
aproximado aproximar VMP00SM 1
de de SPS00 0.999984
17 17 Z 1
ciclones ciclón NCMP000 1
para para SPS00 0.999103
el el DA0MS0 1
Pacífico pacífico NP00000 1
cambiando cambiar VMG0000 1
la el DA0FS0 0.972269
cifra cifra NCFS000 0.799844
a a SPS00 0.996023
21 21 Z 1
-- -- Fg 1
```

**Figura 4.7: Noticias lematizadas**

## 4.5. Fase 4: Análisis del corpus

Para esta etapa del método de solución, se implementó un módulo que realiza un análisis estadístico del *Corpus* de entrenamiento. Este módulo genera las frecuencias de verbos, palabras, gramas, número de noticias analizadas, promedio de oraciones por noticia y el promedio de palabras por oración. Estos resultados son almacenados en archivos de texto generados por cada categoría.

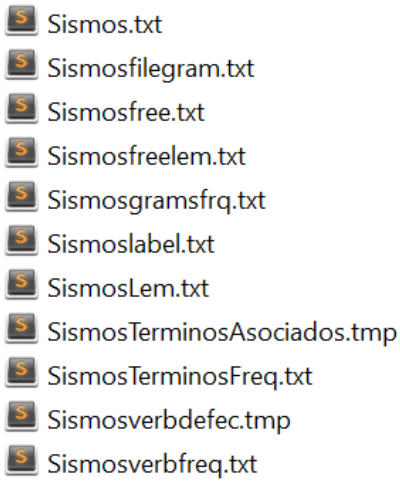



**Figura 4.8: Banco de datos de noticias analizadas.**

Los documentos de nuestro banco de datos contienen:

- **[Nombre de la categoría].txt:** Este documento contiene todas las noticias etiquetadas respecto a su categoría (ver Figura 4.8).
- **[Nombre de la categoría]filegram.txt:** Se almacenan los ngramas.
- **[Nombre de la categoría]free.txt:** Documento generado por el analizador sintáctico Freeling (ver Figura 4.7).
- **[Nombre de la categoría]freelem.txt**: Lista de palabras lematizadas.
- **[Nombre de la categoría]gramsfrq.txt:** Contiene la frecuencia de los gramas.
- **[Nombre de la categoría]label.txt:** Documento que contienen la lista de palabras de la noticia y su etiqueta morfológica, generada por Freeling.
- **[Nombre de la categoría]lem.txt:** Archivo que contiene las noticias lematizadas bajo el analizador sintáctico Freeling, sin stopwords.
- **[Nombre de la categoría]verb.txt:** Contenedor de los verbos y las oraciones donde aparece.
- **[Nombre de la categoría]verbfreq.txt:** Marca la frecuencia de los verbos respecto a su categoría.
- **[Nombre de la categoría]TerminosAsociados.txt:** Contenedor de los términos a fines a cada categoría.
- **[Nombre de la categoría]TerminosFreq.txt:** Contiene los términos frecuentes de cada categoría.
- **[Nombre de la categoría]VerbDefec.txt:** Documento que contiene los verbos defectivos de cada categoría.

## 4.6. Fase 5: Extracción de hechos

Para realizar la extracción de hechos se empleó un método basado en una identificación de patrones empleando una heurística. Se describe a continuación.

### 4.6.1. Heurística

Para la obtención de patrones se generó una heurística conformada por cuatro variantes las cuales fueron desarrolladas a partir del análisis de las noticias que previamente se descargaron. Estas variantes se basan en tomar ciertas palabras para analizar su contexto e identificar apariciones frecuentes. Las palabras tomadas por las variantes se denominan "palabras base" o "pivote" y las definimos como aquellas palabras que están contenidas en hechos y cuyo contexto puede ser predecible porque tiene apariciones muy frecuentes.

La variante uno toma como pivote los verbos frecuentes. Estos verbos son calculados utilizando los verbos que se encuentran en el total de noticias del *Corpus*.

La variante dos tiene como palabra base los verbos meteorológicos (también llamados defectivos) asociados a las categorías previamente definidas en el capítulo Generación de corpus.

La variante tres toma los términos frecuentes de cada categoría. Los términos son contabilizados utilizando el total de noticias encontradas en el corpus.

En la variante cuatro las palabras base son palabras asociadas a cada categoría. Es decir, las palabras con las que es posible conformar un campo semántico, por ejemplo para la categoría de ciclones las palabras base son: lluvia, viento, ciclón, tormentas, temperaturas, probabilidad, precipitaciones, entre otros.

La heurística entonces identifica en la noticia el pivote de acuerdo a la variante aplicada y se analiza su contexto, es decir, se verifican las palabras que se encuentran delante y detrás de éste para identificar el hecho, como se muestra en la figura.

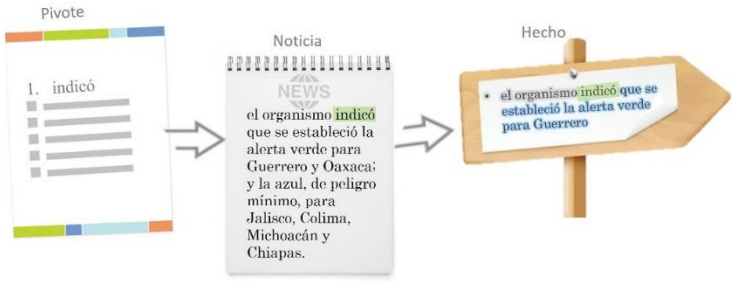



**Figura 4.9: Ejemplo del uso del pivote para iniciar el proceso de extracción de hechos.**

### 4.6.2. Obtención de patrones

Para extraer patrones se creó un algoritmo con dos métodos, el primer método es para identificar oraciones con las etiquetas POS (Apéndice A) de las noticias analizadas y el segundo método extrae los patrones a partir del método de identificación de oraciones.

El método realiza lo siguiente:

1. Los documentos de entrada para este método son: el documento de noticias con etiquetas POS y documento de términos con sus etiquetas.
3. Leer el documento de términos, así como el documento de noticias etiquetadas.
4. Se verifica si la etiqueta leída del documento es “EN” (etiqueta que señala el inicio o fin de la noticia)
5. Entonces se escribe en un nuevo documento llamado “[categoría] párrafo” si no, se lee la palabra siguiente y se compara con el documento de términos.
6. Si el resultado de la comparación es falsa, es decir la palabra que se compara del documento de noticias no existe en el documento de términos, se agrega la etiqueta POS de esa palabra a una cadena para concatenar las etiquetas de palabras que no se encuentren en el documento de términos.
7. Si el resultado de la comparación es verdadera, indicando que si existe la palabra en el documento de términos, se agrega a la cadena término encontrado formando un párrafo de etiquetas que tiene como palabra base el término encontrado.

La siguiente figura muestra un diagrama de flujo del método.

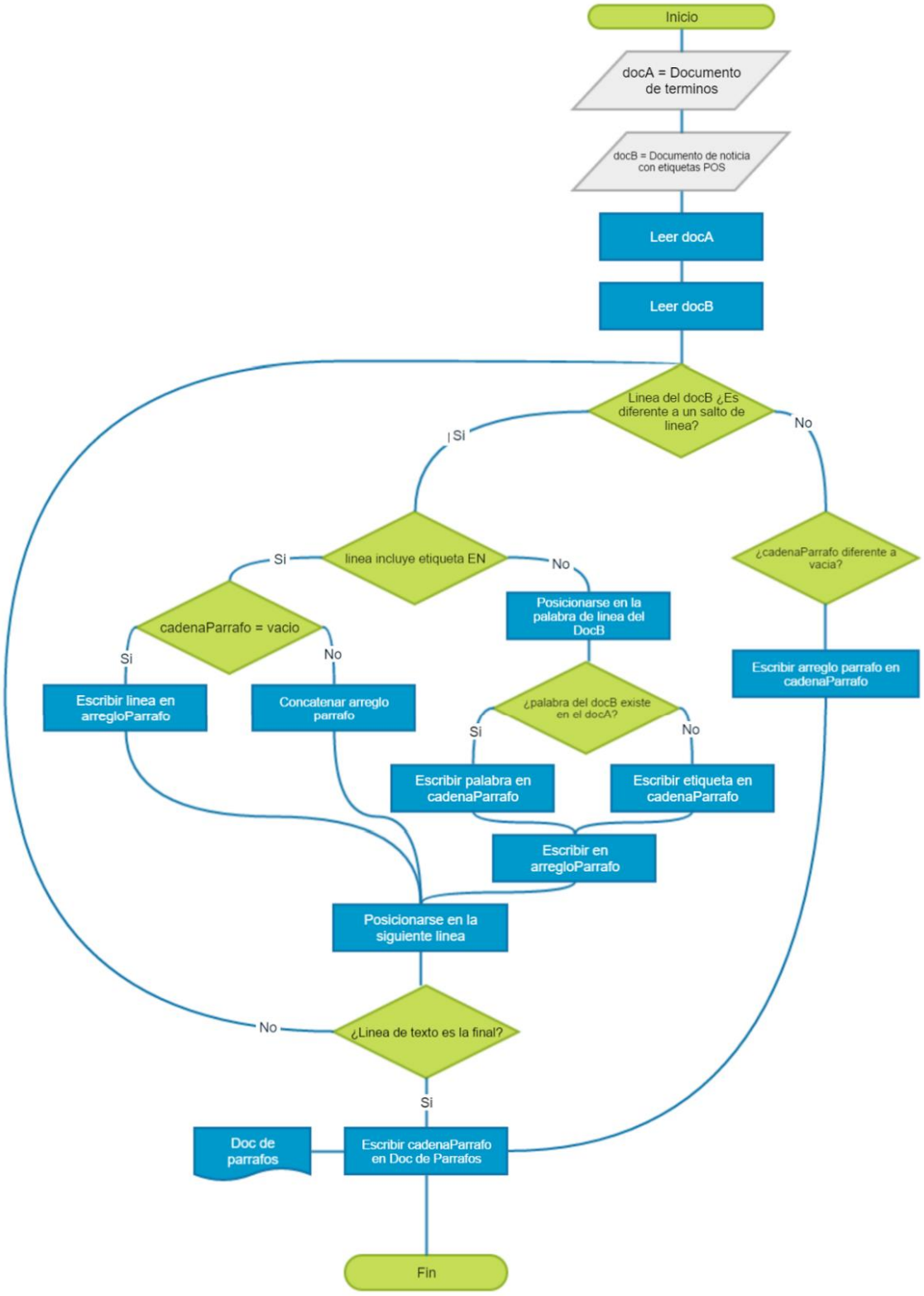



**Figura 4.10: Diagrama de flujo del método obtención de párrafos por etiquetas.**

El método 2 es de Identificación de patrones se ejecuta lo siguiente:

1. Los documentos de entrada son: el documento de oraciones generado del método **identificar oraciones** y el documento de términos. Leer el documento de términos y posteriormente el documento de noticias etiquetadas.
2. Si en el párrafo existe la palabra base, se realizan recorridos a lo largo del texto llevando a cabo lo siguiente: El inicio del párrafo se toma como el inicio del patrón. El fin del primer recorrido será la etiqueta que se encuentra después de la palabra base y se busca el patrón en una estructura de datos donde se almacena los patrones identificados.
   2.1. Si el patrón no existe en la estructura de datos, se almacena y se recorre a la siguiente etiqueta. Se vuelve a realizar la búsqueda y se recorren las etiquetas hasta que se terminan las etiquetas que se encuentran a la derecha de la palabra base.
3. El procedimiento anterior se realiza pero con la diferencia de que el recorrido es hacia la izquierda. Los patrones almacenados en la estructura de datos se escriben en un documento llamado “[Categoria]Patrones”.

   En la siguiente figura se muestra un diagrama de flujo del método.

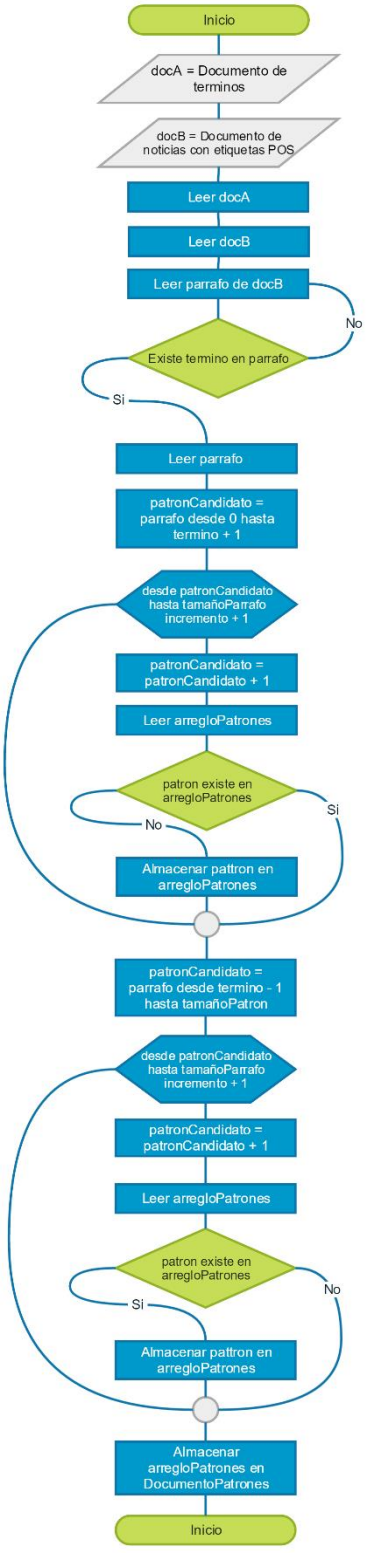



**Figura 4.11: Diagrama de flujo de identificación de patrones.**

El resultado de los patrones encontrados se muestra a continuación. Para la variante 1:

**Con noticias lematizadas** se presentó un total de 2137 patrones de los cuales 1% pertenecen a la categoría de ciclones, 60% helada, 6% inundaciones, 14% precipitación pluvial, 5% a sequía, 4% sismos, 6% tormentas y 4% a tsunamis.

**Con noticias sin lematizar** se extrajeron 2091 patrones con los cuales 2% son de ciclones, 65% de heladas, 11% de inundaciones, 12% precipitación pluvial, 0.4% de sequía, 4% de sismos, 1% de tormenta y 4% de tsunamis.

En la siguiente tabla se comparan las cantidades obtenidas.

**Tabla 4.2: Tabla general de patrones extraídos de la variante 1.**

| **Lematizadas** | | **Sin lematizar** | |
|---|---|---|---|
| Categoría | Patrones | Categoría | Patrones |
| Ciclones | 15 | Ciclones | 33 |
| Heladas | 1292 | Heladas | 1368 |
| Inundación | 122 | Inundación | 234 |
| Precipitación Pluvial | 300 | Precipitación Pluvial | 252 |
| Sequía | 116 | Sequía | 9 |
| Sismos | 93 | Sismos | 92 |
| Tormenta | 119 | Tormenta | 22 |
| Tsunamis | 80 | Tsunamis | 89 |

La siguiente tabla muestra ejemplos de patrones utilizados con sus hechos obtenidos.

**Tabla 4.3: Tabla de ejemplos de patrones con los hechos encontrados de las categorías ciclones y heladas de la variante 1.**

| **Categoría** | **Patrones** | **Hecho** |
|---|---|---|
| Sismos | NC PR afectar DA NC SP DA NC | ▪ tsunami que afectar el noreste de el país<br>▪ sismo que afectar el norte de el pais |
| | NC se haber VM Z NC SP DA NC | ▪ placa se haber mover 95 centímetro hacia el este<br>▪ hora se haber registrar 105 réplica de el temblor |
| Tormenta | DA NP VM SP NC RG nublar SP NC Fc Z SP NC SP NC AQ SP Z NC | ▪ el mesa_central permanecer con cielo medio nublar a nublado con 60/100 de probabilidad de lluvia menor a 25 milímetro<br>▪ el pacífico_norte seguir con cielo medio nublar a nublado con 80/100 de probabilidad de lluvia menor a 25 milímetro |
| | DA NP haber NC RG VM SP NC | ▪ el pacífico_norte haber cielo medio nublar a nublado<br>▪ el pacífico_sur haber cielo medio nublar a nublado |

El número de patrones encontrados en la variante 2 por categoría es el siguiente:

**Con noticias lematizadas** se presentó un total de 893 patrones donde 1% pertenecen a la categoría de ciclones, 62% heladas, 4% inundaciones, 18% precipitación pluvial, 1% a sequía, 1% sismos, 9% tormentas y 4% a tsunamis.

**Con noticias sin lematizar**, se extrajeron 2334 patrones con los cuales 4% son de ciclones, 24% de heladas, 2% de inundaciones, 7% precipitación pluvial, 1% de sequía, 7% de sismos, 54% de tormenta y 2% de tsunamis.

En la siguiente tabla se observan los valores de los porcentajes antes señalados.

**Tabla 4.4: Tabla general de patrones extraídos de la variante 2.**

| **Lematizadas** | | **Sin lematizar** | |
|---|---|---|---|
| Categoría | Patrones | Categoría | Patrones |
| Ciclones | 12 | Ciclones | 93 |
| Heladas | 554 | Heladas | 552 |
| Inundación | 37 | Inundación | 47 |
| Precipitación Pluvial | 159 | Precipitación Pluvial | 152 |
| Sequía | 7 | Sequía | 15 |
| Sismos | 10 | Sismos | 163 |
| Tormenta | 82 | Tormenta | 1267 |
| Tsunamis | 32 | Tsunamis | 45 |

En la siguiente tabla se muestran ejemplos de patrones encontrados y de sus respectivos hechos.

**Tabla 4.5: Tabla de ejemplos de patrones con los hechos encontrados de las categorías ciclones y Sequía de la variante 2.**

| Categoría | Patrones | Hecho |
|---|---|---|
| Ciclones | DA NC CN ubicar SP Z NC SP DA NC SP NP Fc NP | ▪ el fenómeno se ubicar a 335 kilómetro a el sur-suroeste de acapulco , guerrero<br>▪ el meteoro se ubicar a 620 kilómetro a el sur de puerto_ángel , oaxaca |
| | NP VM Z SP NC SP evolucionar SP NC AQ SP DA NC | ▪ océano_pacífico mantener 70/100 de probabilidad de evolucionar a ciclón tropical en el pronóstico<br>▪ guerrero mantener 70/100 de probabilidad de evolucionar a ciclón tropical en el pronóstico |
| Heladas | DA NP CN VM NC RG nublar DA AQ NC SP DA NC Fc NC SP NC | ▪ el península_de_yucatán se tener cielo medio nublar el mayor parte de el día<br>▪ el pacífico_sur se esperar cielo medio nublar el mayor parte de el día |
| | DA NP VM NC despejar DA AQ NC SP DA NC | ▪ la Península_de_Yucatán persistir cielo despejar la mayor parte de el día<br>▪ el Pacífico_Norte dominar cielo despejar la mayor parte de el día |

El número de patrones encontrados en la variante 3 son:

**Con noticias lematizadas** se presentó un total de 2296 patrones de los cuales 48% son de ciclones, 15% heladas, 3% inundaciones, 27% precipitación pluvial, 2% a sequía, 2% sismos, 2% tormentas y 1% a tsunamis.

**Con noticias sin lematizar** se extrajeron 1329 patrones con los cuales 7% son de ciclones, 42% de heladas, 14% de inundaciones, 11% precipitación pluvial, 4% de sequía, 11% de sismos, 9% de tormenta y 2% de tsunamis.

Los valores obtenidos se comparan en la siguiente tabla.

**Tabla 4.6: Tabla general de patrones extraídos de la variante 3.**

| **Lematizadas** | | **Sin lematizar** | |
|---|---|---|---|
| Categoría | Patrones | Categoría | Patrones |
| Ciclones | 1112 | Ciclones | 93 |
| Heladas | 355 | Heladas | 552 |
| Inundación | 59 | Inundación | 182 |
| Precipitación Pluvial | 617 | Precipitación Pluvial | 152 |
| Sequía | 43 | Sequía | 54 |
| Sismos | 49 | Sismos | 145 |
| Tormenta | 35 | Tormenta | 119 |
| Tsunamis | 26 | Tsunamis | 32 |

En la siguiente tabla se muestran ejemplos de patrones encontrados y de sus respectivos hechos.

**Tabla 4.7: Tabla de ejemplos de patrones con los hechos encontrados de las categorías tormenta y sismos de la variante 3.**

| **Categoría** | **Patrones** | **Hecho** |
|---|---|---|
| Tormenta | DA NP VM CS DA temperatura AQ SP DA AO Z NC P0 VM SP NP Fc NP | ▪ la conagua reportar que la temperatura máxima de las últimas 24 horas se registrar en el_gallo , guerrero<br>▪ el meteorológico precisar que la temperatura máxima de las últimas 24 horas se registrar en tapachula , chiapas |
| | DA AO tormenta AQ VA RG SP DA NC SP NP | ▪ la quinta tormenta invernal estar hoy sobre el sur de sonora<br>▪ la quinta tormenta invernal estar hoy sobre el sur de chihuahua |
| Sismos | DA NC SP DA sismo P0 VM SP Z NC | ▪ el epicentro de el sismo se ubicar a 86 kilómetros<br>▪ el hipocentro de el sismo se localizar a 52.2 kilómetros |
| | DA AO tormenta AQ PR CN VM SP NP | ▪ la tercer tormenta invernal que se localizar sobre Nuevo_México<br>▪ la quinta tormenta invernal que se encontrar sobre Baja_California |

La cantidad de patrones extraídos en esta variante para la variante 4 son:

**Con noticias lematizadas** fue un total de 1192 patrones mediante el cual 2% pertenecen a la categoría de ciclones, 33% heladas, 8% inundaciones, 8 precipitación pluvial, 6% a sequía, 8% sismos, 27% tormentas y 9% a tsunamis.

**Con noticias sin lematizar**, se obtuvieron 1496 patrones con los cuales 3% son de ciclones, 29% de heladas, 10% de inundaciones, 18% precipitación pluvial, 6% de sequía, 6% de sismos, 21% de tormenta y 7% de tsunamis.

En la siguiente tabla se muestran las cantidades de las categorías mencionadas.

**Tabla 4.8: Tabla general de patrones extraídos de la variante 4.**

| **Lematizadas** | | **Sin lematizar** | |
|---|---|---|---|
| categoría | patrones | categoría | patrones |
| Ciclones | 27 | Ciclones | 47 |
| Heladas | 392 | Heladas | 434 |
| Inundación | 92 | Inundación | 145 |
| Precipitación Pluvial | 92 | Precipitación Pluvial | 270 |
| Sequía | 71 | Sequía | 89 |
| Sismos | 92 | Sismos | 86 |
| Tormenta | 319 | Tormenta | 316 |
| Tsunamis | 107 | Tsunamis | 109 |

En la siguiente tabla se muestran ejemplos de patrones encontrados y de sus respectivos hechos.

**Tabla 4.9: Tabla de ejemplos de patrones con los hechos encontrados de las categorías Sequía e inundación de la variante 4.**

| **Categoría** | **Patrones** | **Hecho** |
|---|---|---|
| Sequía | DA NC RG afectar VS NP Fc NP | ▪ el estado más afectar ser chihuahua , durango<br>▪ el estado más afectar ser coahuila , nuevo_león |
| | DA AQ sequía PR CN VA VM SP DA NC | ▪ el peor sequía que se haber registrar en el norte<br>▪ el mayor sequía que se haber registrar en el país |
| Inundación | DA NP Z VM lluvias AQ SP NC SP NP | ▪ la onda_tropical 30 ocasionar lluvias intensas en zonas de chiapas<br>▪ la onda_tropical 32 favorecer lluvias intensas con tormentas en chiapas |
| | DA NC SP lluvias AQ SP DA AQ NC | ▪ el pronóstico de lluvias intensas durante los próximos días<br>▪ el pronóstico de lluvias fuertes en los próximos días |

Se filtraron los patrones obtenidos eliminando aquellos que no extraían un hecho correcto. Como ejemplo se tiene el siguiente candidato:

"NP *informar* CS DA NC CN"

Extrae el siguiente texto:

"Miguel_Ángel_Mancera informar que el centro se"

Como puede apreciarse el texto extraído no tiene sentido al estar incompleto, por lo tanto no puede ser un hecho.

Este tipo de patrones son descartados y solo se consideran los que extraen hechos completos. Por ejemplo, el patrón:

"DA NC CN ubicar SP Z NC SP DA NC SP NP"

Extrae el siguiente hecho:

"el fenómeno se ubicar a 335 km a el sur-suroeste de Acapulco".

Este patrón nos da hechos correctos mostrando como sujeto "el fenómeno", el verbo "ubicar" y como predicado/complemento "a 335 km a el sur-suroeste de Acapulco".

El apéndice C muestra ejemplos de los patrones identificados. Los patrones se presentan de acuerdo con las categorías de desastres naturales que se analizan en esta investigación.

### 4.6.3. Identificación de hechos

Para identificar los hechos utilizando los patrones, se elaboró un algoritmo que busca las concordancias de las etiquetas de cada patrón en las etiquetas del documento de noticias etiquetadas.

El algoritmo ejecuta lo siguiente:

1. Los documentos de noticias etiquetadas y el de patrones, se utilizan como documentos de entrada.
2. Se selecciona el patrón.
3. Se recorre y comparan las etiquetas con las etiquetas del documento de noticias etiquetadas.
4. Si la etiqueta del patrón es igual a la etiqueta del documento de noticias etiquetadas, almacenar en una cadena el lema de la etiqueta encontrada. Este proceso se realiza hasta que se encuentran todas las etiquetas del patrón.
5. Si el patrón contiene una palabra base, se concatena la palabra a la cadena de etiquetas almacenadas.

6. Si el patrón no se encuentra en la noticia o se rompe en cualquier etiqueta sin dejar terminar de recorrer el patrón, se vuelve a iniciar el recorrido del el siguiente patrón con las etiquetas a partir de donde no se encontró la etiqueta del documento de noticias etiquetadas. Cada concordancia se escribe en un nuevo documento denominada [Categoría]Hechos.

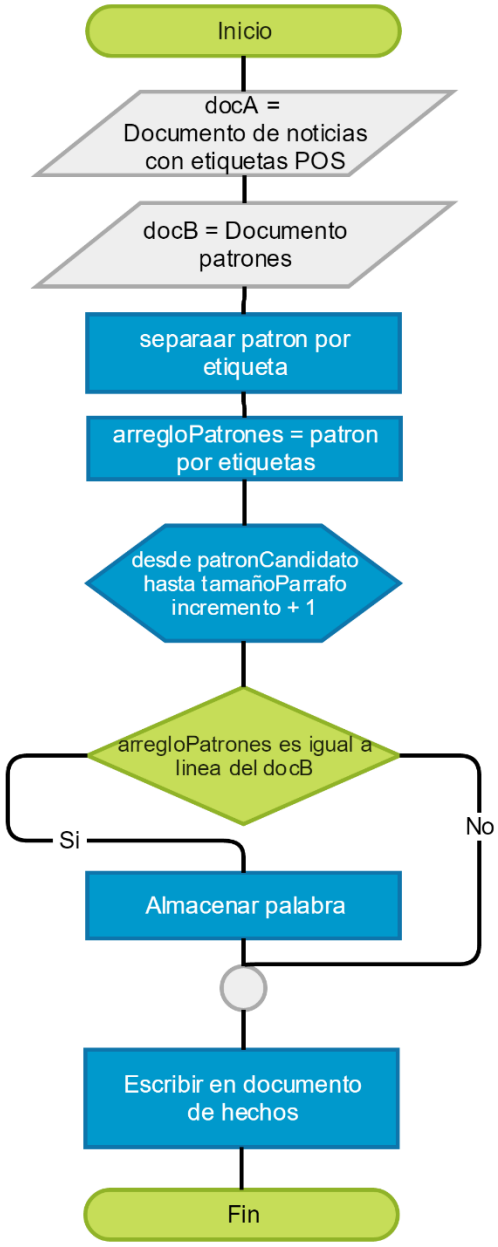



**Figura 4.12: Diagrama de flujo del algoritmo para la extracción de hechos.**

### 4.6.4. Aplicación de reglas para extraer los hechos

Se aplicaron tres reglas para obtener los hechos. Las reglas empleadas fueron las siguientes:

#### 4.6.4.1. Regla 1.

La primera regla que se aplicó a los patrones fue eliminar etiquetas que no puedan iniciar o finalizar un patrón. Algunas de estas etiquetas son:

Punto - FP
Comas - FC
Preposiciones – SP
Verbos – VM
Que – PR
Y – CC
Adjetivo – AQ, entre otros.

Se eliminaron aquellos patrones que solo contenían una o dos etiquetas antes y después de la palabra base. Por ejemplo, estos dos primeros casos son patrones que se eliminan.

**Caso 1:**

Patrón “DA NP Fpa NP Fpt prever NC”

Este patrón es eliminado porque solo contiene una etiqueta a la derecha del pivote. Esto indica que no extrae un hecho completo. El patrón solo obtendrá este texto:

“la comisión_nacional_de_el_agua ( conagua ) prever temperaturas”

Dejando incompleto el predicado del hecho.

**Caso 2:**

Patrón: “RG P0 prever NC AQ SP DA NC”

En el caso de este patrón, aunque contiene dos etiquetas antes del pivote, inicia con la etiqueta RG indicando que es un adverbio y no contiene un sujeto antes. El texto que obtiene este patrón es:

“también se prever temperaturas frías durante la mañana”.

Aunque después del pivote se muestra un predicado correcto, hace falta un sujeto que determine la persona o cosa que realiza la acción.

Estos casos muestran textos incompletos extraídos por el patrón, ya sea del lado del predicado como el caso 1 o del sujeto como el caso 2. Si el patrón contiene dos o menos

etiquetas del lado derecho y una o ninguna etiqueta antes del pivote (lado izquierdo), el patrón será eliminado.

El siguiente caso da un ejemplo que es contemplado como un patrón correcto.

**Caso 3:**

Patrón: DA NC SP NC prever SP DA NP NC VM SP NC VM Fc NC SP NC

El texto extraído es el siguiente:

"el pronóstico por región prever en el pacífico_norte clima despejar a medio nublar , probabilidad de lluvias"

Este patrón no es eliminado ya que extrae un hecho completo mostrando el sujeto "el pronóstico por región", verbo (pivote de este patrón) "prever" y predicado/complemento "en el pacífico_norte clima despejar a medio nublar , probabilidad de lluvias"

#### 4.6.4.2. Regla 2.

Esta regla se aplica al encontrase una conjunción "y" en los hechos candidatos extraídos, para formar dos hechos. Por ejemplo:

"La onda tropical 20 aumentó a 60% su potencial para evolucionar a ciclón tropical en las próximas 48 horas **y** a 90% en su pronóstico de tres a cinco días"

Al aplicar la regla se formarán los siguientes hechos:

**Hecho 1**. La onda tropical 20 aumentó a 60% su potencial para evolucionar a ciclón tropical en las próximas 48 horas

**Hecho 2**. La onda tropical 20 aumentó a 90% en su pronóstico de tres a cinco días.

### 4.6.5. Categorización de patrones

Se tomaron los patrones de etiquetas gramaticales y manualmente se construyeron elementos que permitieran un mejor manejo de los patrones para posteriormente extraer hechos.

Se categorizaron los patrones de acuerdo al verbo encontrado en el hecho. La clasificación está conformado por verbos defectivos, verbos de acción y verbos de habla (ver ejemplos en Apéndice B).

Los ejemplos de elementos designados se encuentran en la Tabla 4.10.

**Tabla 4.10: Ejemplo de elementos encontrados.**

| Elementos | Ejemplos |
|---|---|
| <lugar> | pacífico_norte, mesa_de_el_norte, pacífico_sur, península_de_yucatán. |
| <porcentaje> | 70%, 25%, 9%. |
| <condicion_admosferica> | cielo despejar a medio, cielo despejar, probabilidad de lluvia. |
| <unidad_medida> | 25 milímetro |
| Verbos defectivos | nublar, sostener, granizar. |

Los verbos defectivos, son aquellos que no tienen conjugación completa y son también llamados "verbos meteorológicos" (Elvira, n.d.). Las características de los patrones que contienen un verbo defectivo son similares por ejemplo: el porcentaje, condición atmosférica, o una unidad de medida.

El grupo categorizado por verbos defectivos se muestra en la Figura 4.13.

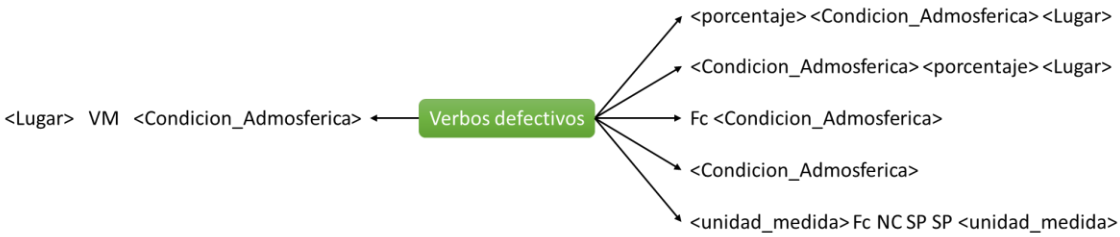



**Figura 4.13: Arquitectura del patrón "Verbos defectivos".**

Un ejemplo del resultado obtenido es:

**Tabla 4.11: Ejemplo de un hecho encontrado con el patrón "verbos defectivos"reconocido.**

| **Elemento** | **Patrón** |
|---|---|
| el península_de_yucatán | <lugar> |
| predominar | VM |
| cielo despejar a medio | <condición_admosferica> |
| nublar | Verbo defectivo |
| , probabilidad de lluvia | Fc <condición_admosferica> |

La siguiente categoría es de verbos de habla. Estos indican acciones comunicativas o expresan creencia, reflexión o emoción y sirven para introducir un discurso (Sidorov, n.d.).

El grupo categorizado de patrones se muestra en la Figura 4.14.

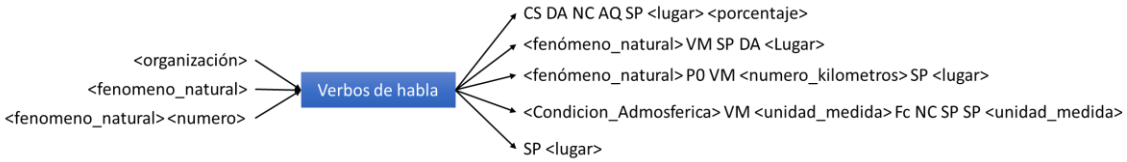



**Figura 4.14: Arquitectura del patrón "Verbos de habla"**

La Tabla 4.12 muestra un ejemplo de un hecho encontrado a partir del patrón "<fenómeno_natural> verbos de habla <fenómeno_natural> P0 VM <número_kilometros> <lugar>".

**Tabla 4.12: Ejemplo de un hecho encontrado con el patrón "verbos de habla" reconocido.**

| Entidad encontrada | Patrón |
|---|---|
| el servicio_meteorológico_nacional | <fenómeno_natural> |
| informar | Verbo de habla |
| el meteoro | <fenómeno_natural> |
| se | P0 |
| ubicar | VM |
| 620 kilómetros | <número_kilometros> |
| a el sur de puerto_ángel | <lugar> |

La categoría "Verbos de acción" (Figura 4.15) son aquellos verbos que incluyen noción de ocurrencia de movimiento de traslado. Es el verbo de donde se percibe un proceso de movimiento o cambio de lugar por un período determinado o específico de tiempo (Talmy, 2000)(Messineo & Klein, 2005).

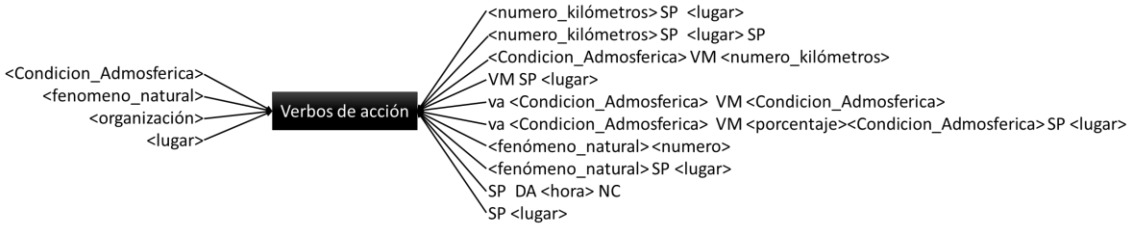



**Figura 4.15: Arquitectura del patrón "Verbos de acción".**

La Tabla 4.13 muestra un ejemplo del hecho extraído con el patrón <fenómeno_natural> verbo de acción <números_kilometros> SP <lugar>.

**Tabla 4.13: Ejemplo de un hecho encontrado con el patrón "verbos acción" reconocido.**

| Entidad encontrada | Patrón |
|---|---|
| el sismo | <fenómeno_natural> |
| ocurrir | verbo de acción |
| 40 kilómetro | <número_kilometros> |
| a | SP |
| sureste de pijijilpan | <lugar> |

## 4.7. Implementación del sistema en un Servicio Web

El servicio web da soporte al método para la extracción de hechos. La finalidad es mostrar al usuario hechos obtenidos de una noticia del periódico seleccionado. La arquitectura se muestra en la Figura 4.16 y el funcionamiento es el siguiente:

La aplicación cliente que accede al servicio web, envía la llamada a través de un controlador. Éste genera la conexión e intercambio de datos entre la aplicación cliente y el servicio web. Una vez que se realiza la conexión, el servicio procesa la noticia y muestra como resultado los hechos extraídos a la aplicación cliente.

El servicio está compuesto por dos módulos que recibe una noticia. En el primer módulo se realiza el preprocesamiento excluyendo los delimitadores que se encuentran en el texto como las etiquetas HTML y etiquetando la noticia insertando una etiqueta al inicio y fin de cada nota.

En el segundo módulo se lleva a cabo la extracción de hechos. Este módulo realiza la extracción de patrones con las noticias preprocesadas. Extrae los hechos conforme los patrones y son almacenados en una base de datos.

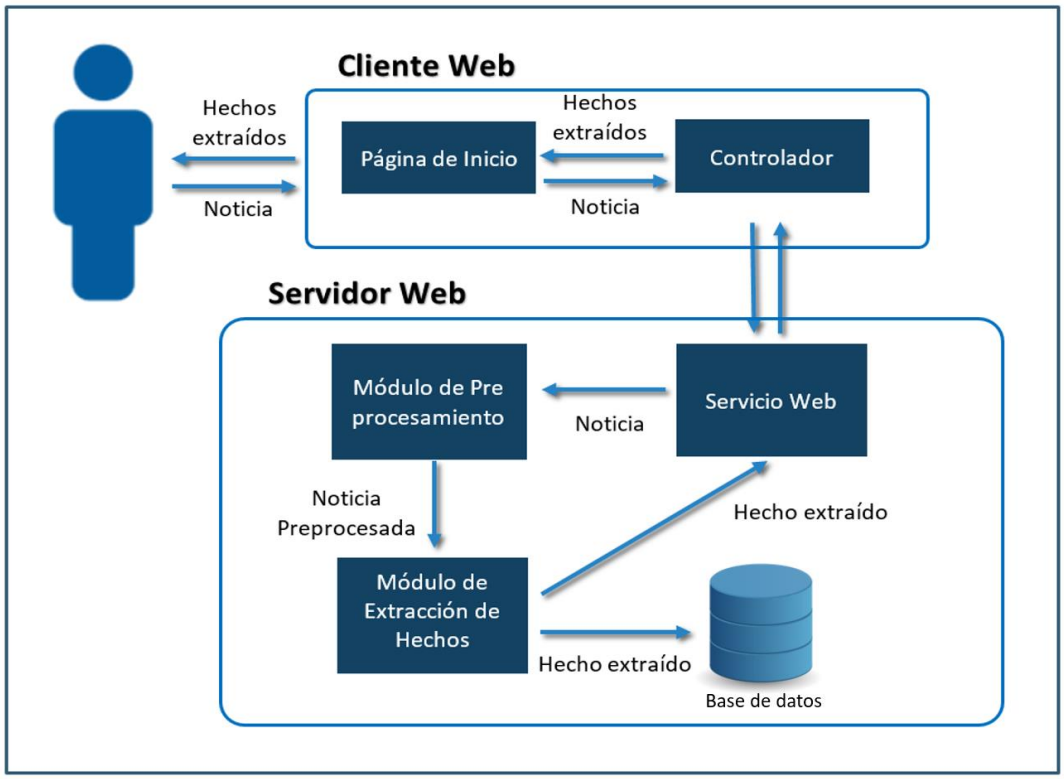



**Figura 4.16: Arquitectura del servicio web.**

Para utilizar el servicio, es necesario obtener la URL (seleccionado con el recuadro rojo en la Figura 4.17) de la noticia así como lo muestra la siguiente figura.


**Figura 4.17: Muestra para obtener la URL de una noticia.**

El usuario debe seleccionar la URL del sitio web del periódico y copiarla para posteriormente insertarla en el campo de texto denominado URL noticia (ver Figura 4.18) del servicio web desarrollado. Presionar el botón “Cargar” para obtener la noticia del periódico.

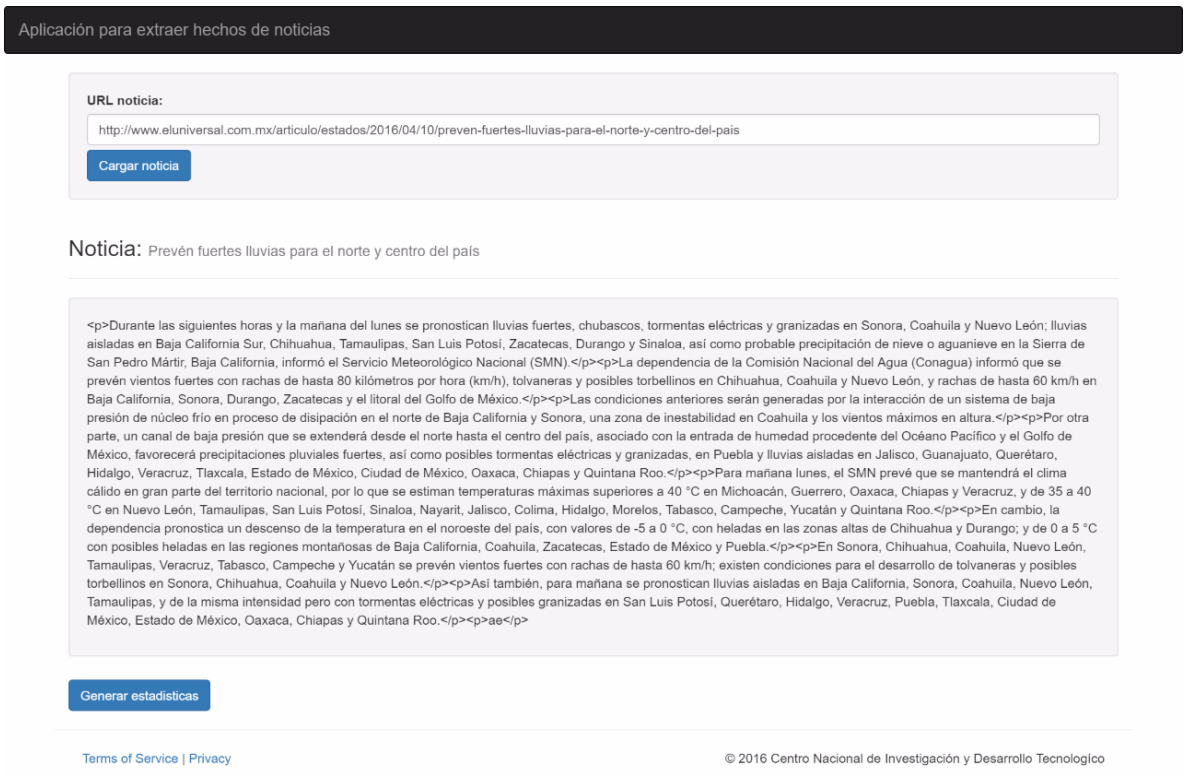



**Figura 4.18: Muestra de campo de texto para la URL de la noticia.**

El sitio muestra el título y el cuerpo de la noticia. Una vez realizado lo anterior se muestra un nuevo botón que permite obtener los hechos como se muestra en la Figura 4.19.

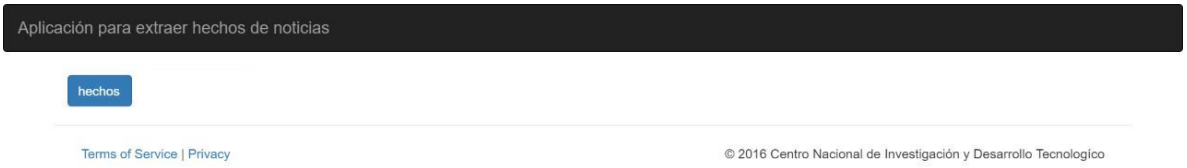



**Figura 4.19: Botón para generar patrones y extraer los hechos.**

Al momento de estar terminado el proceso, el servicio muestra la lista de hechos extraídos en una nueva página (ver Figura 4.20).


**Figura 4.20: Hechos extraídos mostrados en la aplicación.**

# Capítulo 5
# Experimentos y resultados

En este capítulo se presentan los resultados obtenidos de los experimentos realizados sobre la investigación desarrollada.

Para comprobar el grado de rendimiento de la investigación, se aplicaron dos medidas:

- La precisión (P), se mide dividiendo el número de hechos correctos obtenidos por el sistema entre el número total de hechos obtenidos por el sistema.

$$P = \frac{\text{hechos correctos obtenidos por el sistema}}{\text{total de hechos obtenidos por el sistema}}$$

Donde:

  - Los hechos correctos obtenidos por el sistema son los hechos que contienen un sujeto, verbo y predicado.
  - El total de hechos obtenidos por el sistema son aquellos hechos extraídos tanto correctos como incorrectos.

- El cobertura (R) se mide dividiendo el número de hechos correctos obtenidos por el sistema entre el número total de hechos existentes en las noticias (extraídos por un humano).

$$R = \frac{\text{hechos correctos obtenidos por el sistema}}{\text{total de hechos existentes en las noticias}}$$

  - Donde el **total de hechos existentes** en la noticias son los hechos extraídos por el usuario.

Otra métrica utilizada es la Medida-F (F1), es una medida que combina y balancea la precisión y la cobertura. Se obtiene de la siguiente forma $F1 = 2\ \frac{P\,R}{P+R}$:

## 5.1. Corpus de pruebas

El corpus para pruebas se conforma de 80 noticias. Estas son procesadas por el sistema para extraer los hechos utilizando los patrones obtenidos. Para obtener la precisión y cobertura del método se dividió el corpus de prueba en dos grupos. En el primer grupo se lematizaron las noticias y se extrajeron manualmente un total de 5038 hechos. En el segundo grupo las noticias no se lematizaron y se obtuvieron manualmente 5400 hechos. Todos los hechos se almacenaron en un archivo de texto plano para posteriormente ser comparados con los hechos extraídos por el sistema.

## 5.2. Experimento

El experimento se desarrolla extrayendo hechos por medio de los patrones obtenidos en la fase 5 del método de solución. Se aplicaron las cuatro variantes existentes en la heurística. Los resultados se presentan a continuación.

### 5.2.1. Variante 1

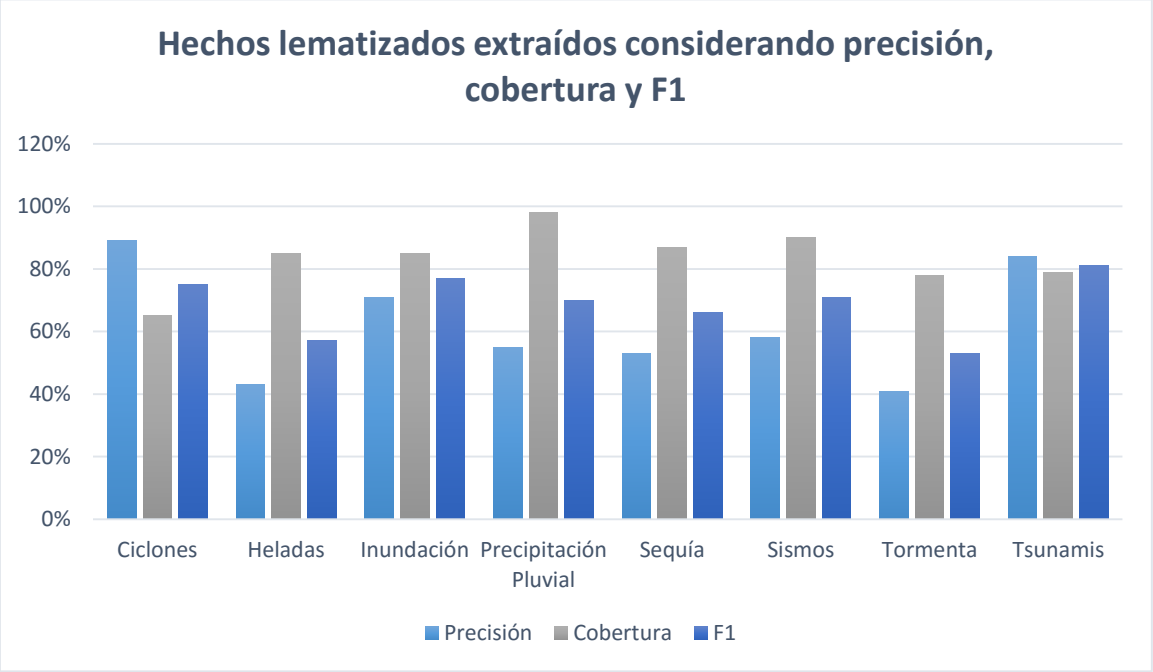
Los resultados para la **variante 1** con noticias lematizadas y sin lematizar se muestran a continuación:

**Con noticias lematizadas** se extrajeron automáticamente 10315 hechos, de los cuales 4764 fueron correctos, logrando alcanzar una precisión de 46%, una cobertura de 86% y una medida F de 60%. La siguiente tabla muestra el resultado de hechos obtenidos por el sistema y por un experto.

**Tabla 5.1: Tabla de hechos lematizados extraídos.**

| Categorías | Hechos obtenidos por el sistema | Hechos correctos obtenidos por el sistema | Hechos correctos obtenidos por un experto |
|---|---|---|---|
| Ciclones | 36 | 32 | 49 |
| Heladas | 7719 | 3286 | 3885 |
| Inundación | 253 | 179 | 211 |
| Precipitación Pluvial | 1356 | 744 | 758 |
| Sequía | 236 | 126 | 145 |
| Sismos | 200 | 116 | 129 |
| Tormenta | 349 | 142 | 182 |
| Tsunamis | 166 | 139 | 176 |

La Gráfica 5.1 muestra una gráfica con el resultado del porcentaje de precisión, cobertura y medida F de la extracción de hechos de noticias lematizadas. Se indica el resultado obtenido por cada categoría, mostrando que la categoría de **ciclones** tuvo mayor precisión con un **89%** ante las otras categorías.



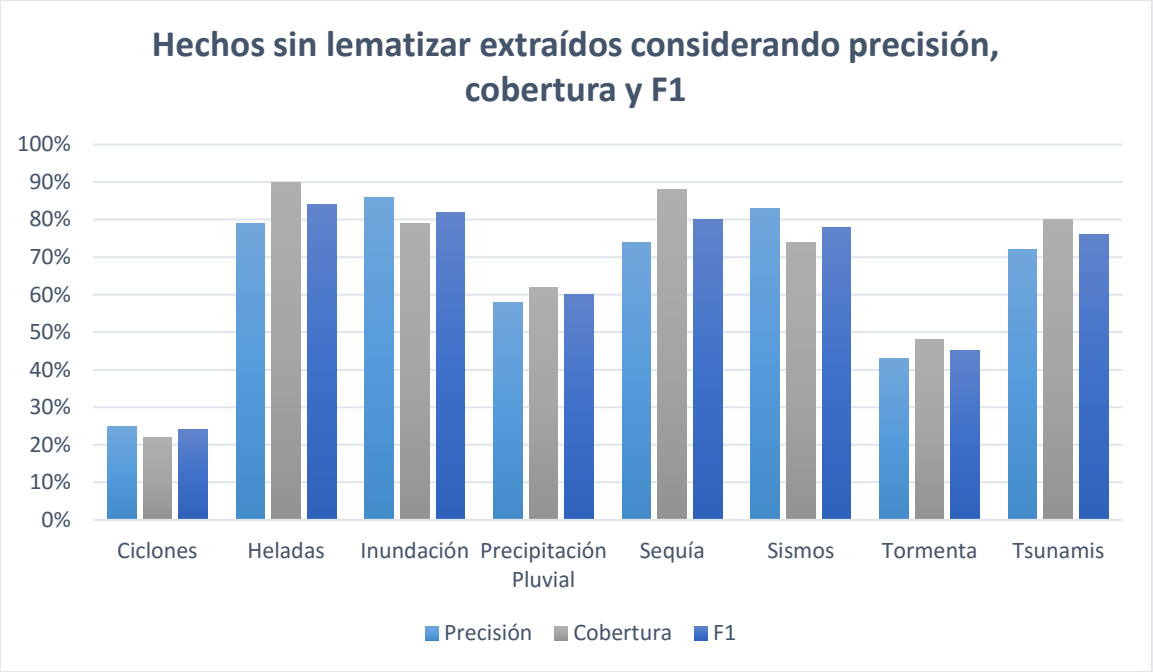

**Gráfica 5.1: Gráfica de hechos lematizados extraídos considerando precisión, cobertura y F1.**

**Con noticias sin lematizar** se obtuvieron 5855 hechos, 3648 correctos presentando una precisión de 62%, cobertura de 63% y una medida F de 63%. La siguiente tabla muestra el resultado de hechos obtenidos por el sistema y por un experto.

**Tabla 5.2: Tabla de hechos sin lematizar extraídos.**

| Categorías | Hechos obtenidos por el sistema | Hechos correctos obtenidos por el sistema | Hechos correctos obtenidos por un experto |
|---|---|---|---|
| Ciclones | 1144 | 286 | 1290 |
| Heladas | 1685 | 1334 | 1485 |
| Inundación | 681 | 587 | 747 |
| Precipitación Pluvial | 1342 | 779 | 1265 |
| Sequía | 164 | 121 | 138 |
| Sismos | 324 | 270 | 365 |
| Tormenta | 349 | 151 | 317 |
| Tsunamis | 166 | 120 | 150 |

La Gráfica 5.1 muestra una gráfica con el resultado del porcentaje de precisión, cobertura y medida F de la extracción de hechos de noticias sin lematizar. La gráfica indica el resultado obtenido por cada categoría, mostrando un puntaje de precision alto en la categoría de **inundaciones** con un **86%** ante las otras categorías.




**Gráfica 5.2: Gráfica de hechos sin lematizar extraídos considerando precisión, cobertura y F1.**

## 5.2.2. Variante 2

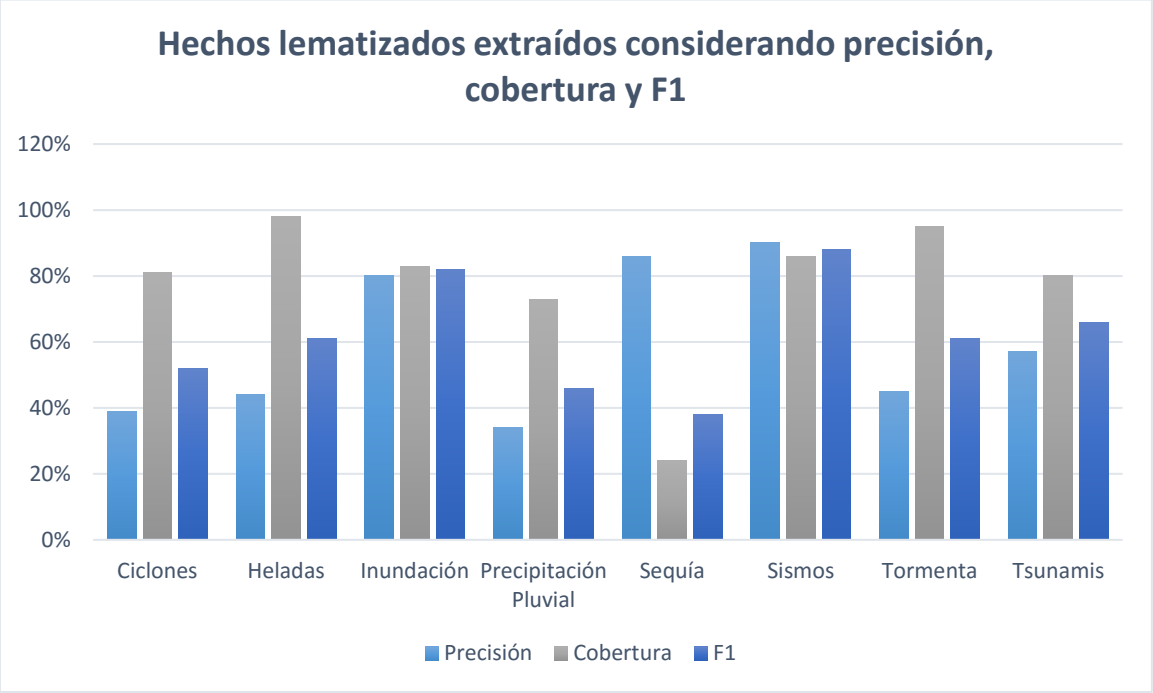
Aplicando la **variante 2** se obtuvieron los siguientes valores:

**Con noticias lematizadas** el sistema obtuvo 5998 hechos, 2601 fueron correctos, mostrando una precisión de 43% con un cobertura de 91% y medida F de 59%. La siguiente tabla muestra el resultado de hechos obtenidos por el sistema y por un experto.

**Tabla 5.3: Tabla de hechos lematizados extraídos.**

| Categorías | Hechos obtenidos por el sistema | Hechos correctos obtenidos por el sistema | Hechos correctos obtenidos por un experto |
|---|---|---|---|
| Ciclones | 44 | 17 | 21 |
| Heladas | 4497 | 1982 | 2028 |
| Inundación | 75 | 60 | 72 |
| Precipitación Pluvial | 990 | 334 | 457 |
| Sequía | 14 | 12 | 49 |
| Sismos | 42 | 38 | 44 |
| Tormenta | 267 | 119 | 125 |
| Tsunamis | 69 | 39 | 49 |

La Gráfica 5.3 muestra el resultado del porcentaje de precisión, cobertura y medida F de la extracción de hechos con noticias lematizadas. Se indica el resultado obtenido por cada categoría, mostrando que la categoría de **sismos** tuvo mayor precisión con un **90%** ante las demás categorías.



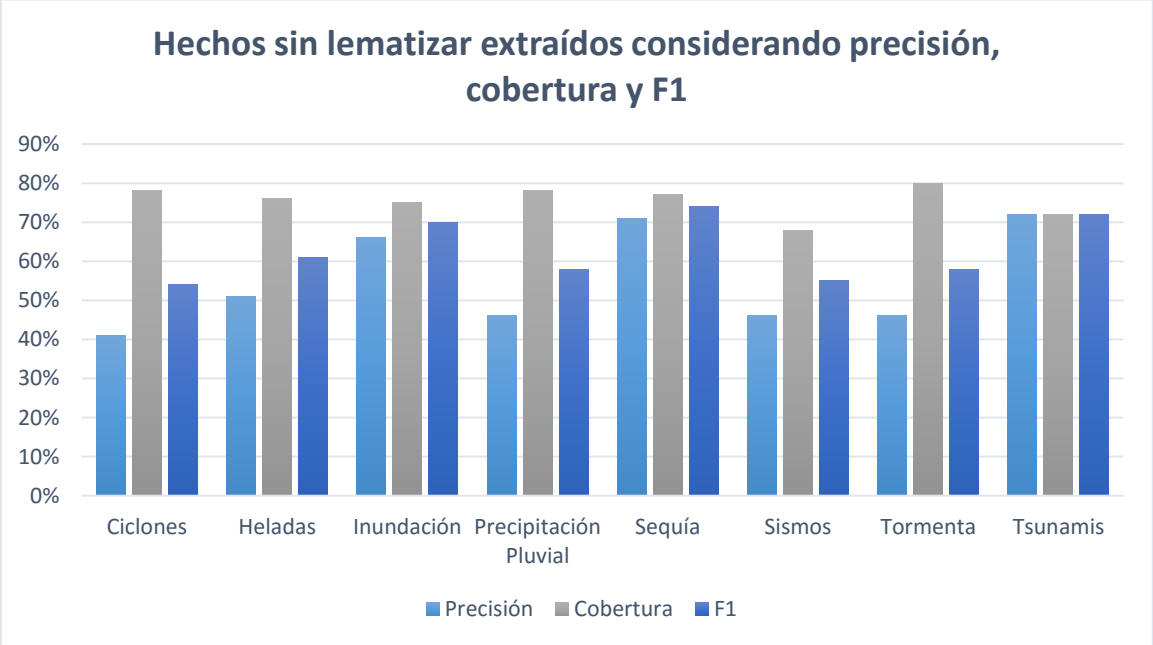

**Gráfica 5.3: Gráfica de hechos lematizados extraídos considerando precisión, cobertura y F1.**

**Con las noticias sin lematizar**, el sistema generó 6377 hechos, 3062 correctos cubriendo una precisión de 48%, un cobertura de 77% y una medida F de 59%. La siguiente tabla muestra el resultado de hechos obtenidos por el sistema y por un experto.

**Tabla 5.4: Tabla de hechos sin lematizar extraídos considerando precisión, cobertura y F1.**

| Categorías | Hechos obtenidos por el sistema | Hechos correctos obtenidos por el sistema | Hechos correctos obtenidos por un experto |
|---|---|---|---|
| Ciclones | 396 | 163 | 209 |
| Heladas | 4493 | 2278 | 2983 |
| Inundación | 32 | 21 | 28 |
| Precipitación Pluvial | 976 | 352 | 449 |
| Sequía | 34 | 24 | 31 |
| Sismos | 28 | 13 | 19 |
| Tormenta | 349 | 161 | 202 |
| Tsunamis | 69 | 50 | 69 |

La Gráfica 5.1 muestra una gráfica con el resultado general del porcentaje de precisión, cobertura y medida F de la extracción de hechos con noticias sin lematizar. Además, indica el resultado obtenido por cada categoría, mostrando que la categoría de **tsunamis** tuvo mayor precisión con un **72%** ante las demás categorías.




**Gráfica 5.4: Gráfica de hechos sin lematizar extraídos considerando precisión, cobertura y F1.**

### 5.2.3. Variantes 3

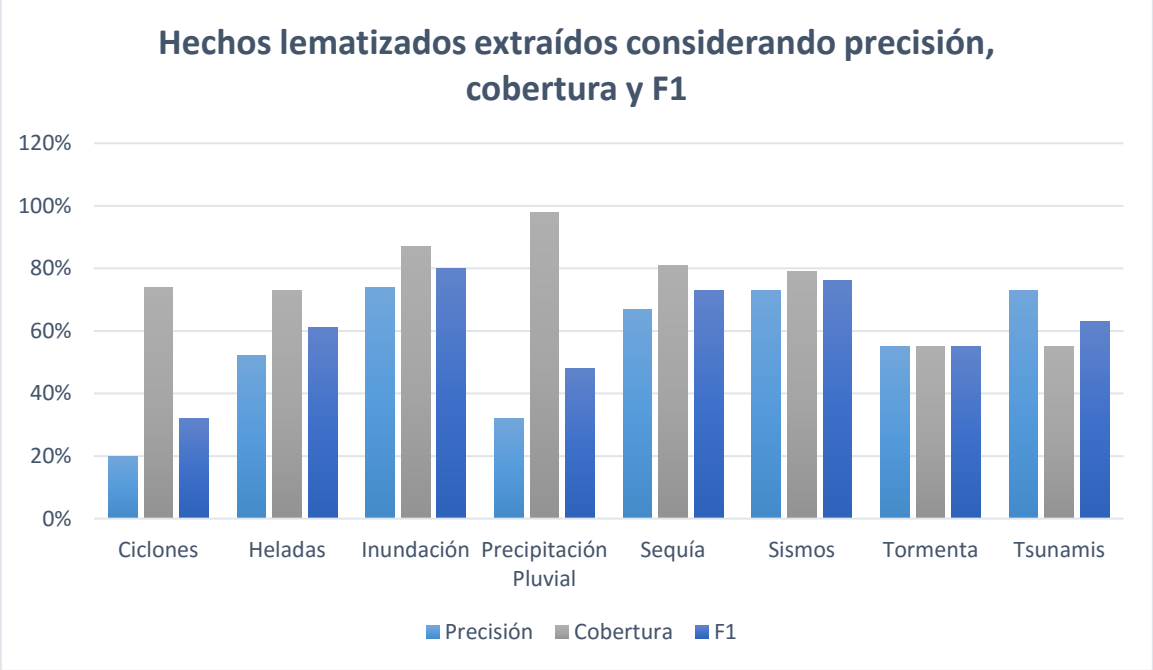
Para la **variante 3** se obtuvieron los siguientes datos:

**Con noticias lematizadas** se generaron 3890 hechos, 1296 fueron obtenidos de forma correcta logrando tener una precisión del 33%, cobertura de 89% y una medida F de 48%. La siguiente tabla muestra el resultado de hechos obtenidos por el sistema y por un experto.

**Tabla 5.5: Tabla de hechos lematizados extraídos considerando precisión, cobertura y F1.**

| Categorías | Hechos obtenidos por el sistema | Hechos correctos obtenidos por el sistema | Hechos correctos obtenidos por un experto |
|---|---|---|---|
| Ciclones | 887 | 179 | 241 |
| Heladas | 5096 | 2637 | 3590 |
| Inundación | 129 | 95 | 109 |
| Precipitación Pluvial | 2616 | 839 | 852 |
| Sequía | 87 | 58 | 72 |
| Sismos | 107 | 78 | 99 |
| Tormenta | 369 | 203 | 367 |
| Tsunamis | 64 | 47 | 85 |

La Gráfica 5.5 muestra una gráfica con el resultado general del porcentaje de precisión, cobertura y medida F de la extracción de hechos con noticias lematizadas. Se muestra el total de hechos extraídos por cada categoría mostrando a **inundación** como la categoría con mayor precisión.



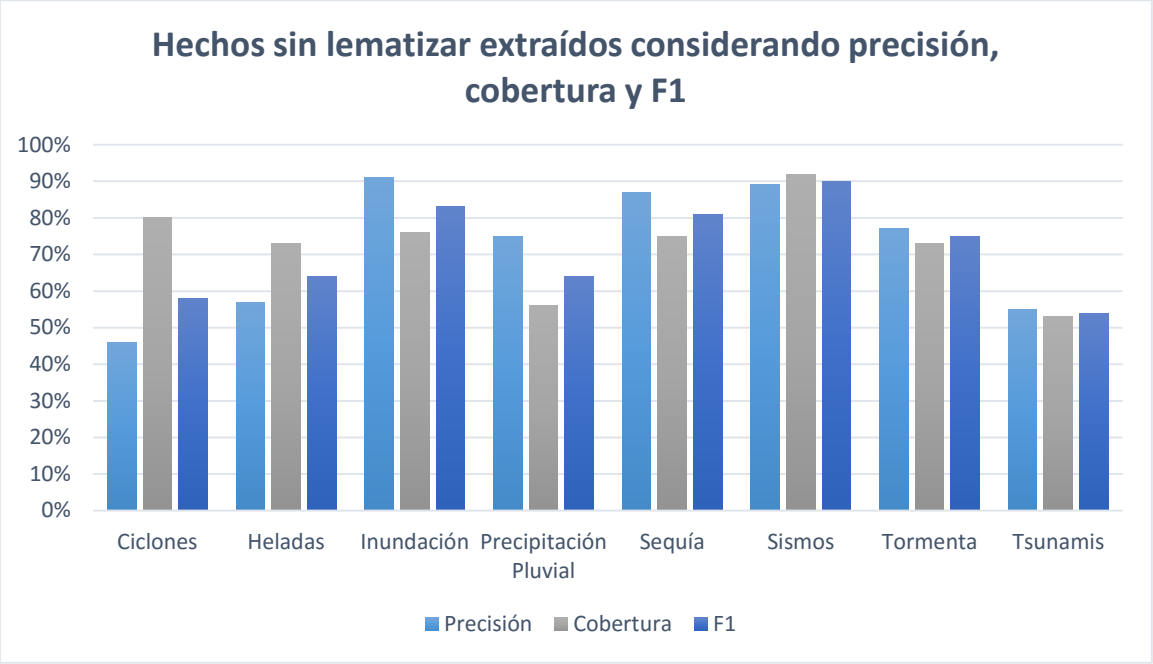

**Gráfica 5.5: Gráfica de hechos lematizados extraídos considerando precisión, cobertura y F1.**

**Con las noticias sin lematizar**, se obtuvieron 1918 hechos, 1085 correctos con precisión de 57% y cobertura del 82% y medida F de 67%. La siguiente tabla muestra el resultado de hechos obtenidos por el sistema y por un experto.

**Tabla 5.6: Tabla de hechos sin lematizar extraídos considerando precisión, cobertura y F1.**

| Categorías | Hechos obtenidos por el sistema | Hechos correctos obtenidos por el sistema | Hechos correctos obtenidos por un experto |
|---|---|---|---|
| Ciclones | 1455 | 671 | 843 |
| Heladas | 5058 | 2860 | 3904 |
| Inundación | 130 | 118 | 156 |
| Precipitación Pluvial | 204 | 152 | 270 |
| Sequía | 129 | 112 | 149 |
| Sismos | 333 | 296 | 322 |
| Tormenta | 809 | 624 | 857 |
| Tsunamis | 384 | 210 | 396 |

La Gráfica 5.6 muestra una gráfica con el resultado general del porcentaje de precisión, cobertura y medida F de la extracción de hechos con noticias sin lematizar. La Tabla 5.6 indica el resultado obtenido por cada categoría, mostrando que la categoría de **inundación** tuvo mayor precisión con un **91%** ante las demás categorías.




**Gráfica 5.6: Gráfica de hechos sin lematizar extraídos considerando precisión, cobertura y F1.**

### 5.2.4. Variante 4

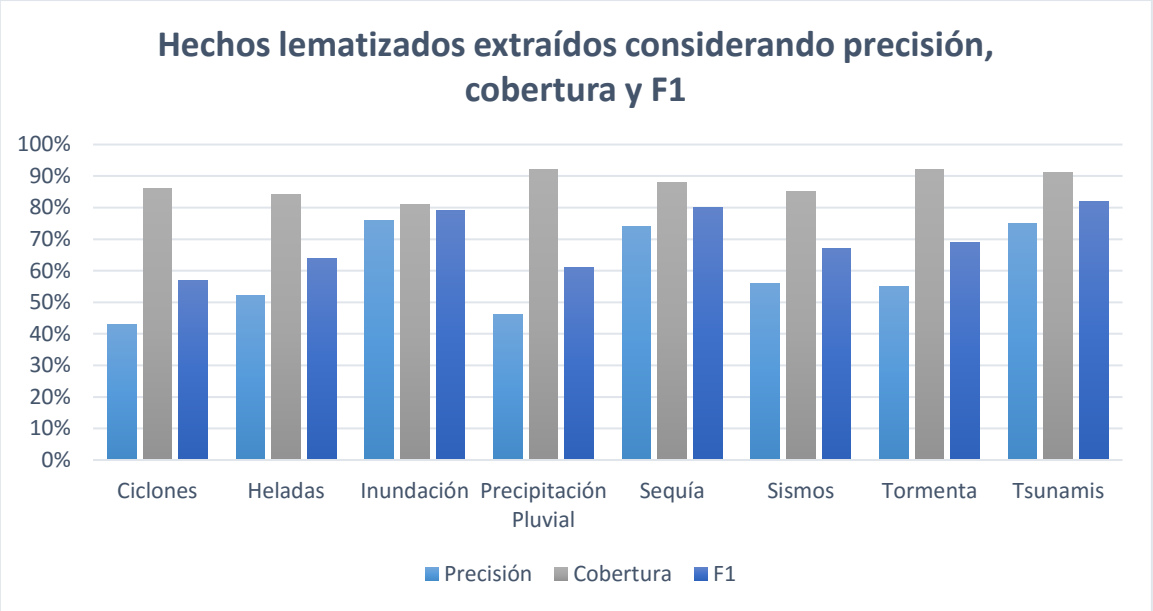
La **cuarta variante** muestra un total de hechos como se describe a continuación:

**Con noticias lematizadas** el total de hechos generados por el sistema fue de 5220, 2661 fueron correctos con una precisión de 51%, 91% de cobertura y 82% de medida F. La siguiente tabla muestra el resultado de hechos obtenidos por el sistema y por un experto.

**Tabla 5.7: Tabla de hechos lematizados extraídos considerando precisión, cobertura y F1.**

| Categorías | Hechos obtenidos por el sistema | Hechos correctos obtenidos por el sistema | Hechos correctos obtenidos por un experto |
|---|---|---|---|
| Ciclones | 310 | 132 | 154 |
| Heladas | 5690 | 2932 | 3501 |
| Inundación | 186 | 142 | 175 |
| Precipitación Pluvial | 3207 | 1470 | 1590 |
| Sequía | 143 | 106 | 121 |
| Sismos | 198 | 111 | 131 |
| Tormenta | 927 | 514 | 561 |
| Tsunamis | 249 | 186 | 204 |

La Gráfica 5.7 muestra una gráfica con el resultado general del porcentaje de precisión, cobertura y medida F de la extracción de hechos con noticias lematizadas. Además, la gráfica indica el resultado obtenido por cada categoría, mostrando que la categoría de **Inundación** tuvo mayor precisión con un **76%** ante las demás categorías.



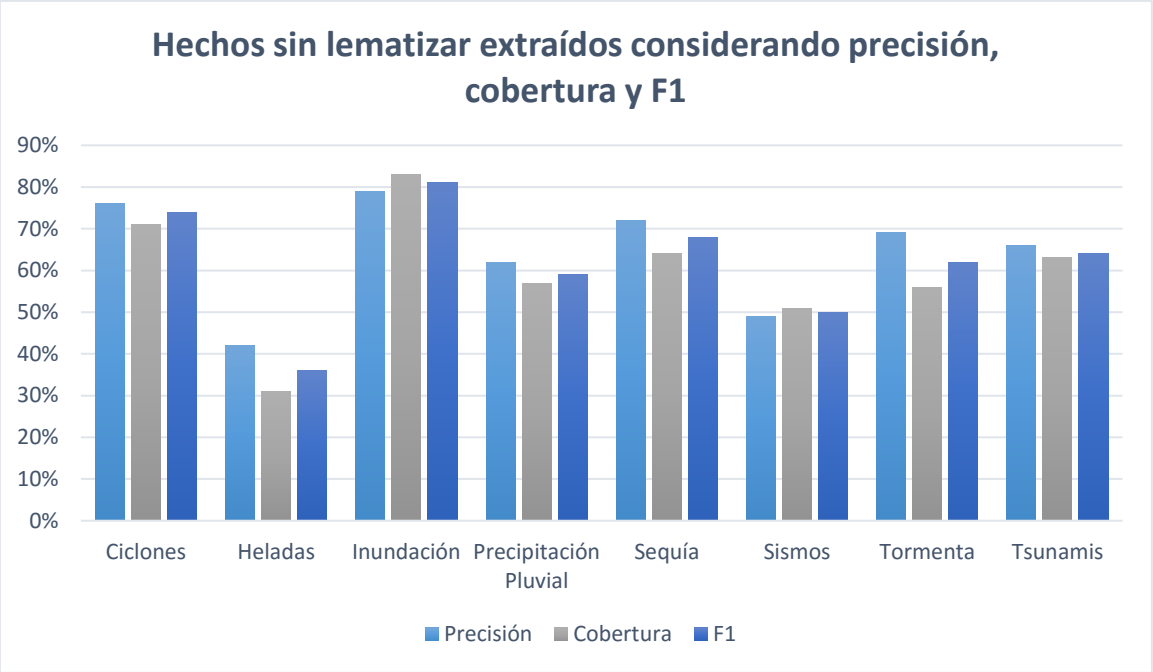

**Gráfica 5.7: Gráfica de hechos lematizados extraídos considerando precisión, cobertura y F1.**

**Con noticias sin lematizar** dieron como resultado 1950 hechos extraídos por el sistema, 1652 fueron correctos con una precisión de 85%, un cobertura de 83% y una medida F de 84%. La Tabla 5.8 muestra el resultado de hechos obtenidos por el sistema y por un experto.

**Tabla 5.8: Tabla de hechos sin lematizar extraídos considerando precisión, cobertura y F1.**

| Categorías | Hechos obtenidos por el sistema | Hechos correctos obtenidos por el sistema | Hechos correctos obtenidos por un experto |
|---|---|---|---|
| Ciclones | 1109 | 842 | 1182 |
| Heladas | 390 | 164 | 527 |
| Inundación | 508 | 403 | 486 |
| Precipitación Pluvial | 803 | 496 | 869 |
| Sequía | 730 | 523 | 811 |
| Sismos | 333 | 164 | 322 |
| Tormenta | 399 | 275 | 493 |
| Tsunamis | 258 | 169 | 269 |

La Gráfica 5.8 muestra una gráfica con el resultado general del porcentaje de precisión, cobertura y medida F de la extracción de hechos con noticias sin lematizar. Además muestra un precisión alta en la categoría inundación.




**Gráfica 5.8: Gráfica de hechos sin lematizar extraídos considerando precisión, cobertura y F1.**

## 5.2.5. Resultado general por variante

La Tabla 5.9 muestra el resultado general del porcentaje de precisión, cobertura y medida F de la extracción de hechos contrastando cada una de las variantes con noticias lematizadas y sin lematizar.

Se muestra que en promedio se obtuvo mejor precisión en noticias sin lematizar pero la puntuación de la cobertura fue mejor con noticias lematizadas.

**Tabla 5.9: Promedio de porcentajes de los resultados más altos de los hechos obtenidos por variantes.**

| Variantes | Lematizado | | | Sin Lematizar | | |
|---|---|---|---|---|---|---|
| | Precisión | Cobertura | F1 | Precisión | Cobertura | F1 |
| 1 | 46% | 86% | 60% | 62% | 63% | 63% |
| 2 | 43% | 91% | 59% | 48% | 77% | 59% |
| 3 | 44% | 76% | 56% | 59% | 73% | 65% |
| 4 | 51% | 87% | 64% | 67% | 61% | 64% |

### 5.2.6. Resultado general por categoría

Los hechos extraídos con los patrones arrojaron mejores resultados de precisión en las categorías ciclones, inundación y sismos. En las categorías heladas, sismos y tsunamis tuvieron mejores resultados en cobertura.

En la Tabla 5.10 se muestran los porcentajes más altos obtenidos en la aplicación de las diferentes variantes. Las celdas verdes corresponden a resultados obtenidos con noticias lematizadas y las celdas blancas con noticias no lematizadas.

**Tabla 5.10: Promedio de porcentajes de los resultados más altos de los hechos obtenidos por categoría.**

| Variante | Categoría | Precisión | Cobertura |
|---|---|---|---|
| 1 | Ciclones | **89%** | 65% |
| 1 | Heladas | 79% | **90%** |
| 3 | Inundación | **91%** | 76% |
| 3 | Precipitación Pluvial | 75% | 56% |
| 3 | Sequía | 87% | 75% |
| 2 | Sismos | **90%** | **86%** |
| 4 | Tormenta | 69% | 56% |
| 1 | Tsunamis | 84% | **79%** |
| | **Resultado** | **83%** | **73%** |

# Capítulo 6
# Conclusiones

A pesar de las múltiples investigaciones que se han realizado sobre el tratamiento de texto relacionado a la extracción de información semántica (extracción de hechos), en el idioma español no ha habido una investigación amplia.

La tendencia general de los trabajos realizados hasta el momento, es la utilización de un análisis basado en árboles sintácticos. En esta investigación se propuso un método diferente basado en estadísticas aplicadas a periódicos en línea de mayor circulación en el país para extraer hechos de noticias relacionadas con desastres naturales. Es bien sabido que esta información puede ser utilizada en tareas como la traducción automática, sistemas de diálogo interactivos, análisis de opiniones, entre otros.

En suma, se realizaron las siguientes actividades:

- Se generó un corpus de noticias de desastres naturales.
- Se desarrolló un módulo de descarga de noticias, el cual puede realizar la descarga de varios sitios web por medio de una URL.
- Se desarrolló un algoritmo que permite la extracción de patrones utilizando etiquetas gramaticales del analizador sintáctico Freeling.
- Se aplicó un análisis estadístico a las noticias para extraer los hechos.
- Se realizó un servicio web para identificar los hechos de la noticia que se ingresa a través de una dirección URL.

Se observó que el implementar el método propuesto para la extracción de hechos es una tarea ardua pero eficaz ya que presenta un valor de precisión del 82% con una cobertura de 92%, que pudo considerarse aceptable como un primer acercamiento a la identificación de hechos a través de métodos estadísticos.

Las noticias de desastres naturales se dividieron en ocho categorías que son ciclones, heladas, inundaciones, precipitación pluvial, sismos, Sequía, tormenta y tsunamis. Se implementaron diversas variantes de análisis estadístico a cada categoría de noticias, y se observó que algunas variantes reportan mejores resultados de precisión y cobertura en ciertas categorías, pero en otras tienen un comportamiento diferente. Por ejemplo, basándose en términos frecuentes se obtuvo la mayor precisión en la categoría de inundación con un 91% lematizando noticias. Sin embargo, esta variante resulta ser una de las que peores resultados arroja para la categoría ciclones con un resultado de **20%** de precisión, también con noticias lematizadas.

## Trabajos futuros

- Continuar con la descarga de noticias para aumentar el corpus y mejorar la precision y cobertura.
- Realizar una categorización de noticias automatiza para facilitar la tarea.
- Realizar una supervisión detallada en la identificación de patrones para obtener mejores resultados con la extracción de hechos.
- Mejorar el servicio web para que pueda ser utilizado en sistemas de investigaciones de interesados.

# Apéndice A:

## Etiquetas morfológicas Eagle empleadas por Freeling

El analizador morfológico Freeling codifica la información morfológica en etiquetas basadas etiquetas propuestas por el grupo "Expert Advisory Group on Language Engineering Standards" **(EAGLES).** Este etiqueta las variables donde cada carácter corresponde a una característica morfológica. Primer carácter de la etiqueta es siempre la categoría (POS). La categoría determina la longitud de la etiqueta y la interpretación de cada carácter.

A continuación presentamos las etiquetas que el analizador morfológico utiliza para el castellano en formato de tabla y algunos ejemplos de cada categoría.

### Adjetivos

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **A**: Adjetivo |
| 1 | Tipo | **O**: Ordinal; **Q**: Calificativo; **P**: Posesivo |
| 2 | Grado | **S**: Superlativo; **V**: Evaluativo |
| 3 | Genero | **F**: Femenino; **M**: Masculino; **C**: Común |
| 4 | Número | **S**: Singular; **P**: Plural ; **N**: Invariable |
| 5 | Possessorpers | **1**: 1ª Persona; **2**: 2 ª Persona; **3**: 3 ª Persona |
| 6 | Possessornum | **S**: Singular; **P**: Plural ; **N**: Invariable |

El lema de los adjetivos siempre es la forma masculina singular (*bonito*) o la forma singular si el adjetivo es de género común (*alegre*). Para los adjetivos invariables, es decir, aquellos que tanto para el singular como para el plural presentan la misma forma, el lema coincide con la forma.

### Ejemplos:

| Forma | Lema | Etiqueta |
|---|---|---|
| alegres | alegre | AQ0CP00 |
| alegre | alegre | AQ0CS00 |
| bonitas | bonito | AQ0FP00 |
| bonita | bonito | AQ0FS00 |
| bonitos | bonito | AQ0MP00 |
| bonito | bonito | AQ0MS00 |
| quemada | quemado | AQ0FS0P |

## Adverbios

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **R**: Adverbio |
| 1 | Tipo | **N**: Negativo; **G**: General |

Para los adverbios, ahora mismo, tan sólo indicamos que es de tipo general. La etiqueta final es RG000 y sirve tanto para los adverbios como para las locuciones adverbiales.

### Ejemplos:

| Forma | Lema | Etiqueta |
|---|---|---|
| despacio | despacio | RG000 |
| ahora | ahora | RG000 |
| siempre | siempre | RG000 |
| hábilmente | hábil | RG000 |
| posteriormente | posterior | RG000 |
| a_cuatro_patas | a_cuatro_patas | RG000 |
| a_granel | a_granel | RG000 |

## Determinantes

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **D**: Determinante |
| 1 | Tipo | **A**: Artículo; **D**: Demostrativo; **I**: Indefinido; **P**: Posesivo; **T**: Interrogativo; **E**: Exclamativo |
| 2 | Persona | **1**: 1ª Persona; **2**: 2 ª Persona; **3**: 3 ª Persona |
| 3 | Género | **F**: Femenino; **M**: Masculino; **C**: Común |
| 4 | Número | **S**: Singular; **P**: Plural; **N**: Invariable |
| 5 | Possessornum | **S**: Singular; **P**: Plural; **N**: Invariable |

**Ejemplos:**

- Determinantes Demostrativos

| Forma | Lema | Etiqueta |
|---|---|---|
| aquel | aquel | DD3MS00 |
| aquella | aquel | DD3FS00 |
| aquellas | aquel | DD3FP00 |
| aquellos | aquel | DD3MP00 |
| esa | ese | DD3FS00 |
| esas | ese | DD3FP00 |
| ese | ese | DD3MS00 |
| esos | ese | DD3MP00 |
| esta | este | DD3FS00 |
| estas | este | DD3FP00 |
| este | este | DD3MS00 |
| estos | este | DD3MP00 |

- Determinantes Posesivos

| Forma | Lema | Etiqueta |
|---|---|---|
| mi | mi | DP3CS01 |
| mis | mi | DP3CP01 |
| tu | tu | DP3CS02 |
| tus | tu | DP3CP02 |
| su | su | DP3CS00 |
| sus | su | DP3CP00 |
| nuestra | nuestro | DP3FS04 |
| nuestras | nuestro | DP3FP04 |
| nuestro | nuestro | DP3MS04 |
| nuestros | nuestro | DP3MP04 |

| | | |
|---|---|---|
| vuestra | vuestro | DP3FS05 |
| vuestras | vuestro | DP3FP05 |
| vuestro | vuestro | DP3MS05 |
| vuestros | vuestro | DP3MP05 |
| suya | suyo | DP3FS00 |
| suyas | suyo | DP3FP00 |
| suyo | suyo | DP3MS00 |
| suyos | suyo | DP3MP00 |

- Determinantes Interrogativos

| Forma | Lema | Etiqueta |
|---|---|---|
| cuánta | cuánto | DT3FS00 |
| cuántas | cuánto | DT3FP00 |
| cuánto | cuánto | DT3MS00 |
| cuántos | cuánto | DT3MP00 |
| qué | qué | DT3CN00 |

- Determinantes Exclamativos

| Forma | Lema | Etiqueta |
|---|---|---|
| qué | qué | DE3CN00 |

- Determinantes Indefinidos

| Forma | Lema | Etiqueta |
|---|---|---|
| alguna | alguno | DI3FS00 |
| algunas | alguno | DI3FP00 |
| alguno | alguno | DI3MS00 |
| algún | alguno | DI3MS00 |
| algunos | alguno | DI3MP00 |
| bastante | bastante | DI3CS00 |
| bastantes | bastante | DI3CP00 |
| cada | cada | DI3CS00 |
| ninguna | ninguno | DI3FS00 |
| ningunas | ninguno | DI3FP00 |
| ninguno | ninguno | DI3MS00 |
| ningún | ninguno | DI3MS00 |
| ningunos | ninguno | DI3MP00 |
| otra | otro | DI3FS00 |
| otras | otro | DI3FP00 |
| otro | otro | DI3MS00 |

| | | |
|---|---|---|
| otros | otro | DI3MP00 |
| sendas | sendos | DI3FP00 |
| sendos | sendos | DI3MP00 |
| tantas | tanto | DI3FP00 |
| tanta | tanto | DI3FS00 |
| tantos | tanto | DI3MP00 |
| tanto | tanto | DI3MS00 |
| todas | todo | DI3FP00 |
| toda | todo | DI3FS00 |
| todos | todo | DI3MP00 |
| todo | todo | DI3MS00 |
| unas | un | DI3FP00 |
| una | un | DI3FS00 |
| unos | un | DI3MP00 |
| un | un | DI3MS00 |
| varias | varias | DI3FP00 |
| varios | varios | DI3MP00 |

## Nombres

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **N**: Sustantivo |
| 1 | Tipo | **C**: Común; **P**: Adecuada |
| 2 | Genero | **F**: Femenino; **M**: Masculino; **C**: Común |
| 3 | Número | **S**: Singular; **P**: Plural; **N**: Invariable |
| 4 | Clasificación Semántica | **S**: Persona; **G**: Ubicación; **O**: Organización; **V**: Otro |
| 5 | Nesubclass | No Utilizado |
| 6 | Grado | **V** : Evaluativa |

Los nombres tienen como lema la forma singular, tanto si es de género femenino como masculino o neutro. Para los nombres invariables, es decir, aquellos que tanto para el singular como para el plural presentan la misma forma (*tesis*), el lema coincide con la forma.

### Ejemplos:

| Forma | Lema | Etiqueta |
|---|---|---|
| chico | chico | NCMS000 |
| chicos | chico | NCMP000 |
| chica | chica | NCFS000 |
| chicas | chica | NCFP000 |
| oyente | oyente | NCCS000 |

| | | |
|---|---|---|
| oyentes | oyente | NCCP000 |
| cortapapeles | cortapapeles | NCMN000 |
| tesis | tesis | NCFN000 |
| Antonio | antonio | NP00000 |

## Verbos

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **V**: Verbo |
| 1 | Tipo | **M**: Principal; **A**: Auxiliar; **S**: Semiauxiliar |
| 2 | Modo | **I**: Indicativo; **S**: Subjuntivo; **M**: Imperativo; **P**: Participio; **G**: Gerundio; **N**: Infinitivo |
| 3 | Tiempo | **P**: Presente; **I**: Imperfecta; **F**: Futuro; **S**: Pasado; **C**: Condicional |
| 4 | Persona | **1**: 1ª Persona; **2**: 2 ª Persona; **3**: 3 ª Persona |
| 5 | Número | **S** : Singular ; **P** : Plural |
| 6 | Genero | **F** : Femenino ; **M** : Masculino ; **C** : Común |

**Ejemplo:**

El lema del verbo siempre es el infinitivo. El atributo de **Género** tan sólo afecta a los participios, para el resto de formas este atributo no se especifica (0).

| Forma | Lema | Etiqueta |
|---|---|---|
| cantada | cantar | VMP00S***F*** |
| cantadas | cantar | VMP00P***F*** |
| cantado | cantar | VMP00S***M*** |
| cantados | cantar | VMP00P***M*** |

## Pronombres

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **P**: Pronombre |
| 1 | Tipo | **D**: Demostrativa; **E**: Exclamativa; **I**: Indefinida; **P**: Personal; **R**: Relativa; **T**: Interrogativa |
| 2 | Persona | **1**:1; **2**:2; **3**:3 |
| 3 | Gen | **F**: Femenino; **M**: Masculino; **C**: Común |
| 4 | Una | **S**: Singular; **P**: Plural; **N**: Invariable |
| 5 | Caso | **N**: Nominativo; **A**: Acusativo; **D**: Dativo; **O**: Oblicua |
| 6 | Politeness | **P**: Sí |

**Ejemplo:**

Pronombres personales

- Los pronombres personales tienen como lema la forma singular: *yo, tú y él*.

| Forma | Lema | Etiqueta |
|---|---|---|
| yo | yo | PP1CSN00 |
| me | yo | PP1CS000 |
| mí | yo | PP1CSO00 |
| nos | yo | PP1CP000 |
| nosotras | yo | PP1FP000 |
| nosotros | yo | PP1MP000 |
| conmigo | yo | PP1CSO00 |
| te | tú | PP2CS000 |
| ti | tu | PP2CSO00 |
| tú | tú | PP2CSN00 |
| os | tú | PP2CP000 |
| usted | tú | PP2CS00P |
| ustedes | tú | PP2CP00P |
| vos | tú | PP3CS00P |
| vosotras | tú | PP2FP000 |
| vosotros | tú | PP2MP000 |
| contigo | tú | PP2CNO00 |
| él | él | PP3MS000 |
| ella | él | PP3FS000 |
| ellas | él | PP3FP000 |
| ello | él | PP3CS000 |
| ellos | él | PP3MP000 |
| la | él | PP3FSA00 |
| las | él | PP3FPA00 |
| lo | él | PP3MSA00 |
| lo | él | PP3CNA00 |
| los | él | PP3MPA00 |
| le | él | PP3CSD00 |
| les | él | PP3CPD00 |
| se | él | PP3CN000 |
| sí | él | PP3CNO00 |
| consigo | él | PP3CNO00 |

- Pronombres demostrativos

| Forma | Lema | Etiqueta |
|---|---|---|
| aquéllas | aquél | PD3FP000 |
| aquélla | aquél | PD3FS000 |
| aquéllos | aquél | PD3MP000 |
| aquél | aquél | PD3MS000 |
| aquellas | aquel | PD3FP000 |
| aquella | aquel | PD3FS000 |
| aquellos | aquel | PD3MP000 |
| aquel | aquel | PD3MS000 |
| aquello | aquello | PD3CS000 |
| ésas | ése | PD3FP000 |
| ésa | ése | PD3FS000 |
| esas | ese | PD3FP000 |
| esa | ese | PD3FS000 |
| esos | ese | PD3MP000 |
| ese | ese | PD3MS000 |
| ésos | ése | PD3MP000 |
| ése | ése | PD3MS000 |
| eso | eso | PD3CS000 |
| esotra | esotro | PD3FS000 |
| esotro | esotro | PD3MS000 |
| esta | este | PD3FS000 |
| éstas | éste | PD3FP000 |
| ésta | éste | PD3FS000 |
| estas | este | PD3FP000 |
| esta | este | PD3FS000 |
| estos | este | PD3MP000 |
| este | este | PD3MS000 |
| éstos | éste | PD3MP000 |
| éste | éste | PD3MS000 |
| esto | esto | PD3CS000 |
| estotra | estotro | PD3FS000 |
| estotro | estotro | PD3MS000 |

- Pronombres posesivos

| Forma | Lema | Etiqueta |
|---|---|---|
| mía | mío | PX3FS010 |
| mías | mío | PX3FP010 |
| mío | mío | PX3MS010 |

| | | |
|---|---|---|
| míos | mío | PX3MP010 |
| nuestra | nuestro | PX3FS040 |
| nuestras | nuestro | PX3FP040 |
| nuestro | nuestro | PX3MS040 |
| nuestros | nuestro | PX3MP040 |
| suya | suyo | PX3FS000 |
| suyas | suyo | PX3FP000 |
| suyo | suyo | PX3MS000 |
| suyos | suyo | PX3MP000 |
| tuya | tuyo | PX3FS020 |
| tuyas | tuyo | PX3FP020 |
| tuyo | tuyo | PX3MS020 |
| tuyos | tuyo | PX3MP020 |
| vuestra | vuestro | PX3FS050 |
| vuestras | vuestro | PX3FP050 |
| vuestro | vuestro | PX3MS050 |
| vuestros | vuestro | PX3MP050 |

- Pronombres indefinidos

| Forma | Lema | Etiqueta |
|---|---|---|
| algo | algo | PI3CN000 |
| alguien | alguien | PI3CN000 |
| alguna | alguno | PI3FS000 |
| algunas | alguno | PI3FP000 |
| alguno | alguno | PI3MS000 |
| algunos | alguno | PI3MP000 |
| cualesquiera | cualquiera | PI3CP000 |
| cualquiera | cualquiera | PI3CS000 |
| demás | demás | PI3CP000 |
| misma | mismo | PI3FS000 |
| mismas | mismo | PI3FP000 |
| mismo | mismo | PI3MS000 |
| mismos | mismo | PI3MP000 |
| mucha | mucho | PI3FS000 |
| muchas | mucho | PI3FP000 |
| mucho | mucho | PI3MS000 |
| muchos | mucho | PI3MP000 |
| nada | nada | PI3CN000 |
| nadie | nadie | PI3CN000 |
| ninguna | ninguno | PI3FS000 |
| ningunas | ninguno | PI3FP000 |
| ninguno | ninguno | PI3MS000 |

| | | |
|---|---|---|
| ningunos | ninguno | PI3MP000 |
| otra | otro | PI3FS000 |
| otras | otro | PI3FP000 |
| otro | otro | PI3MS000 |
| otros | otro | PI3MP000 |
| poca | poco | PI3FS000 |
| pocas | poco | PI3FP000 |
| poco | poco | PI3MS000 |
| pocos | poco | PI3MP000 |
| quienquier | quienquiera | PI3CS000 |
| quienesquiera | quienquiera | PI3CP000 |
| quienquiera | quienquiera | PI3CS000 |
| tanta | tanto | PI3FS000 |
| tantas | tanto | PI3FP000 |
| tanto | tanto | PI3MS000 |
| tantos | tanto | PI3MP000 |
| toda | todo | PI3FS000 |
| todas | todo | PI3FP000 |
| todo | todo | PI3MS000 |
| todos | todo | PI3MP000 |
| última | último | PI3FS000 |
| últimas | último | PI3FP000 |
| último | último | PI3MS000 |
| últimos | último | PI3MP000 |
| una | uno | PI3FS000 |
| unas | uno | PI3FP000 |
| uno | uno | PI3MS000 |
| unos | uno | PI3MP000 |
| varias | varios | PI3FP000 |
| varios | varios | PI3MP000 |

- Pronombres interrogativos

| Forma | Lema | Etiqueta |
|---|---|---|
| adónde | adónde | PT000000 |
| cómo | cómo | PT000000 |
| cuál | cuál | PT3CS000 |
| cuáles | cuál | PT3CP000 |
| cuándo | cuándo | PT000000 |
| cuánta | cuánto | PT3FS000 |
| cuántas | cuánto | PT3FP000 |
| cuánto | cuánto | PT3MS000 |
| cuántos | cuánto | PT3MP000 |

| | | |
|---|---|---|
| dónde | dónde | PT000000 |
| qué | qué | PT3CN000 |
| quién | quién | PT3CS000 |
| quiénes | quién | PT3CP000 |

- Pronombres relativos

| Forma | Lema | Etiqueta |
|---|---|---|
| como | como | PR000000 |
| donde | donde | PR000000 |
| cuando | cuando | PR000000 |
| cual | cual | PR3CS000 |
| cuales | cual | PR3CP000 |
| cuanta | cuanto | PR3FS000 |
| cuantas | cuanto | PR3FP000 |
| cuantos | cuanto | PR3MP000 |
| cuya | cuyo | PR3FS000 |
| cuyas | cuyo | PR3FP000 |
| cuyo | cuyo | PR3MS000 |
| cuyos | cuyo | PR3MP000 |
| que | que | PR3CN000 |
| quien | quien | PR3CS000 |
| quienes | quien | PR3CP000 |

## Conjunciones

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **C**: Conjunción |
| 1 | Tipo | **C**: Coordinación; **S**: Subordinando |

**Ejemplo:**

- Conjunción Coordinada

| Forma | Lema | Etiqueta |
|---|---|---|
| e | e | CC00 |
| empero | empero | CC00 |
| mas | mas | CC00 |
| ni | ni | CC00 |
| o | o | CC00 |
| ora | ora | CC00 |

| | | |
|---|---|---|
| pero | pero | CC00 |
| sino | sino | CC00 |
| siquiera | siquiera | CC00 |
| u | u | CC00 |
| y | y | CC00 |

- Conjunción Subordinada

| Forma | Lema | Etiqueta |
|---|---|---|
| aunque | aunque | CS00 |
| como | como | CS00 |
| conque | conque | CS00 |
| cuando | cuando | CS00 |
| donde | donde | CS00 |
| entonces | entonces | CS00 |
| ergo | ergo | CS00 |
| incluso | incluso | CS00 |
| luego | luego | CS00 |
| mientras | mientras | CS00 |
| porque | porque | CS00 |
| pues | pues | CS00 |
| que | que | CS00 |
| sea | sea | CS00 |
| si | si | CS00 |
| ya | ya | CS00 |

## Numerales

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **Z**: número |
| 1 | Tipo | **d**: partitivo ; **m**: moneda; **p**: porcentaje; **U**: unidad |

**Ejemplo:**

- Numerales Cardinales

| Forma | Lema | Etiqueta |
|---|---|---|
| catorce | catorce | MCCP00 |
| cien | cien | MCCP00 |
| cinco | cinco | MCCP00 |
| cincuenta | cincuenta | MCCP00 |

| | | |
|---|---|---|
| cuatro | cuatro | MCCP00 |
| cuatrocientas | cuatrocientos | MCFP00 |
| cuatrocientos | cuatrocientos | MCMP00 |
| diez | diez | MCCP00 |
| doce | doce | MCCP00 |
| dos | dos | MCCP00 |
| una | uno | MCFS00 |
| unas | uno | MCFP00 |
| uno | uno | MCFS00 |
| unos | uno | MCMP00 |

- Numerales Ordinales

| Forma | Lema | Etiqueta |
|---|---|---|
| primer | primero | MOMS00 |
| primera | primero | MOFS00 |
| primeras | primero | MOFP00 |
| primero | primero | MOMS00 |
| primeros | primero | MOMP00 |
| segundas | segundo | MOFP00 |
| segunda | segundo | MOFS00 |
| segundos | segundo | MOMP00 |
| segundo | segundo | MOMS00 |
| tercer | tercero | MOMS00 |
| terceras | tercero | MOFP00 |
| tercera | tercero | MOFS00 |
| terceros | tercero | MOMP00 |
| tercero | tercero | MOMS00 |
| últimas | último | MOFP00 |
| última | último | MOFS00 |
| últimos | último | MOMP00 |
| último | último | MOMS00 |

# Fecha

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **W**: Fecha |

## Interjecciones

| Posición | Atributos | Valores |
|---|---|---|
| 0 | categoría | I: Interjección |

**Ejemplo:**

| Forma | Lema | Etiqueta |
|---|---|---|
| ah | ah | I |
| eh | eh | I |
| ejem | ejem | I |
| ele | ele | I |

## Preposiciones

| Posición | Atributos | Valores |
|---|---|---|
| 0 | Categoría | **S**: Adposición |
| 1 | Tipo | **P**: Preposición |

**Ejemplos:**

| Forma | Lema | Etiqueta |
|---|---|---|
| al | al | SPCMS |
| del | del | SPCMS |
| a | a | SPS00 |
| ante | ante | SPS00 |
| bajo | bajo | SPS00 |
| cabe | cabe | SPS00 |
| con | con | SPS00 |

## Signos de puntuación

| Etiqueta | Atributos |
|---|---|
| Fd | **Pos**: Puntuación; **Tipo**: Dos Puntos |
| Fc | **Pos**: Puntuación; **Tipo**: Coma |
| Flt | **Pos**: Puntuación; **Tipo**: Llaves; **Clase de Puntuación**: Cierre |
| Fla | **Pos**: Puntuación; **Tipo**: Llaves; **Clase de Puntuación**: Apertura |
| Fs | **Pos**: Puntuación; **Tipo**: Puntos Suspensivos |
| Fat | **Pos**: Puntuación; **Tipo**: Exclamación; |

| | |
|---|---|
| | **Clase de Puntuación**: Cierre |
| Faa | **Pos**: Puntuación; **Tipo**: Exclamación;<br>**Clase de Puntuación**: Apertura |
| Fg | **Pos**: Puntuación; **Tipo**: Guion |
| Fz | **Pos**: Puntuación; **Tipo**: Otro |
| Fpt | **Pos**: Puntuación; **Tipo**: Paréntesis; **Clase de Puntuación**: Cierre |
| Fpa | **Pos**: Puntuación; **Tipo**: Paréntesis; **Clase de Puntuación**: Apertura |
| Ft | **Pos**: Puntuación; **Tipo**: Porcentaje |
| Fp | **Pos**: Puntuación; **Tipo**: Punto |
| Fit | **Pos**: Puntuación; **Tipo**: Signo De Interrogación;<br>**Clase de Puntuación**: Cierre |
| Fia | **Pos**: Puntuación; **Tipo**: Signo De Interrogación;<br>**Clase de Puntuación**: Apertura |
| Fe | **Pos**: Puntuación; **Tipo**: Comillas Inglesas |
| Frc | **Pos**: Puntuación; **Tipo**: Comillas Latinas;<br>**Clase de Puntuación**: Cierre |
| Fra | **Pos**: Puntuación; **Tipo**: Comillas Latinas;<br>**Clase de Puntuación**: Apertura |
| Fx | **Pos**: Puntuación; **Tipo**: Punto Y Coma |
| Fh | **Pos**: Puntuación; **Tipo**: Diagonal |
| Fct | **Pos**: Puntuación; **Tipo**: Corchete; C**lase de Puntuación**: Cierre |
| Fca | **Pos**: Puntuación; **Tipo**: Corchete; **Clase de Puntuación**: Apertura |

**Ejemplos:**

| Forma | Lema | Etiqueta |
|---|---|---|
| ¡ | ¡ | Faa |
| ! | ! | Fat |
| , | , | Fc |
| [ | [ | Fca |
| ] | ] | Fct |
| : | : | Fd |
| " | " | Fe |
| - | - | Fg |
| / | / | Fh |
| ¿ | ¿ | Fia |
| ? | ? | Fit |
| { | { | Fla |
| } | } | Flt |
| . | . | Fp |
| ( | ( | Fpa |
| ) | ) | Fpt |
| ... | ... | Fs |
| % | % | Ft |

# Apéndice B:
## Ejemplo de hechos extraídos

En este apartado se muestran ejemplos de hechos extraídos con los patrones obtenidos divididos por categorías de verbos. En el punto 3.6.5 de este documento se muestra la categorización de patrones la cual se basó en las categorías de verbos.

### Ejemplo con verbos defectivos

| Verbos Defectivos |
|---|
| el pacífico_sur se anunciar cielo despejar a medio nublar con 80100 de probabilidad de lluvia menor a 25 milímetro en chiapas |
| el mesa_central se tener cielo despejar a medio nublar con 20100 de probabilidad de lluvia menor a 25 milímetro en Guanajuato |
| el península_de_yucatán predominar cielo nublar con probabilidad de tormenta fuerte de 60100 en campeche y quintana_roo |
| el mesa_de_el_norte haber cielo nublar con potencial de tormenta fuerte de 60100 en chihuahua y Durango |
| el pacífico_norte haber cielo nublar a medio nublar , probabilidad de lluvia |
| el mesa_de_el_norte prevalecer ambiente nublar a medio nublar con probabilidad de lluvia |
| el mesa_de_el_norte prevalecer ambiente nublar a medio nublar con probabilidad de lluvia |
| el ciclón presentar viento máximo sostener de 100 kilómetro  h con racha de hasta 120 kilómetro  h |

## Ejemplo con verbos de habla

| Verbos de habla |
|---|
| la secretaría_de_turismo informar que la ocupación hotelera en baja_california_sur ser de 49_% y que actualmente haber 30_mil turistas , de los que 26_mil |
| el servicio_meteorológico_nacional informar que el sistema se ubicar sobre la península_de_yucatán |
| el ieepo informar que la determinación se acordar con el instituto_estatal_de_protección_civil |
| el servicio_meteorológico_nacional informar que el meteoro se ubicar a 620 kilómetros a el sur de puerto_Ángel , oaxaca |
| el smn informar que el meteoro se localizar a 320 kilómetros a el oeste de punta_eugenia , bcs |
| el smn informar que el sistema ubicar a 465 kilómetros a el sur de puerto_Ángel |
| el servicio_meteorológico_nacional informar que el sistema ubicar a 465 kilómetros a el sur de puerto_Ángel |
| el ciclón presentar viento máximo sostener de 100 kiló metro h , racha de hasta 120 kilómetro h y movimiento hacia el oeste-noroeste |
| el fenómeno presentar viento máximo sostener de 110 kilómetro hr , racha de hasta 140 kilómetro hr y desplazamiento hacia el noroeste |
| protección_civil estatal indicar que el bajo presión tener potencial ciclónico de 80100 en el próximo 48 hora |
| el meteorológico informar que en el último 24 hora se registrar el temperatura máximo en arriaga |
| el conagua puntualizar que en el último 24 hora se registrar el temperatura máximo en campeche |
| el smn reportar que en el último 24 hora se registrar el temperatura máximo en ejido_nuevo_león |
| el conagua señalar que en el último 24 hora se registrar el temperatura máximo en choix |
| el frente frío 41 se extender desde el norte de el golfo_de_méxico |
| el frente frío 26 se extender sobre el oriente de el golfo_de_méxico |
| canal de bajo presión se extender desde el norte hasta el centro de el territorio |
| canal de bajo presión se extender desde el norte hasta el centro de el país |
| el sistema frontal se extender sobre el norte de chihuahua |
| el sistema frontal se extender desde el costa de estados_unidos |
| el frente frío número 37 se extender en el norte y noreste de el país , para continuar su trayectoria hacia el sureste |
| el frente frío número 43 se extender desde el noreste de estados_unidos |
| el frente frío número 35 se extender desde el noreste de estados_unidos |

| |
|---|
| el servicio_sismológico_nacional indicar que en 14 año más de 50 sismo haber rebasar el 6.0 grado en el escala |
| el movimiento tectónico tener uno magnitud de 6.1 |
| el movimiento telúrico tener uno magnitud de 6.1 |
| el movimiento telúrico se registrar ininterrumpidamente a_partir_de el madrugada |
| el sistema de monitoreo de el volcán popocatépetl registrar 12 exhalación de bajo intensidad |
| el sistema de monitoreo de el volcán popocatépetl registrar 110 exhalación de bajo intensidad |
| titular de la Sedesol , informar que con el fin de atender a las 2,350 comunidades más afectar por la falta de agua |
| la Conagua reportó que la temperatura máxima de las últimas 24 horas se registró en El_Gallo , Guerrero |
| El SMN reportó que la temperatura máxima durante las últimas 24 horas se sintió en Huaquechula , Puebla |
| la Conagua reportó que la temperatura máxima de las últimas 24 horas se registró en El_Gallo , Guerrero , con 45 grados |
| El Meteorológico precisó que la temperatura máxima de las últimas 24 horas se registró en Tapachula , Chiapas , con 34.3 grados |
| El SMN reportó que la temperatura máxima durante las últimas 24 horas se sintió en Huaquechula , Puebla , con 40 grados |

## Ejemplo con verbos de acción

| verbos de accion |
|---|
| el sismo de magnitud 6.8 en el escala de richter que sacudir el zona norte de el costa |
| el instituto_de_geofísica_de_la_universidad_nacional_autónoma_de_méxico precisar que el temblor ocurrir a 40 kilómetro a el suroeste de pijijiapan |
| el instituto_de_geodinámica_de_atenas decir que el sismo ocurrir a 89 kilómetro a el sur de salónica |
| el epicentro de el movimiento telúrico se ubicar a 95 kilómetro de el norteño ciudad de iquique |
| el sismo de 1985 en el ciudad de méxico se trabajar más en el campo de el investigación |
| el temblor ocurrir a 40 kilómetro a el suroeste de pijijiapan |
| el sismo ocurrir a 40 kilómetro a el sureste de pijijilpan |
| el sismo ocurrir a 89 kilómetro a el sur de salónica |
| el sismo ocurrió a 41 kilómetros de profundidad en el mar de las Molucas |
| epicentro se localizó a 75 kilómetros a el suroeste de Panguna , en la isla de Bougainville |
| epicentro se localizó a 259 kilómetros a el suroeste de Tomatlán , en el estado de Jalisco |
| la masa de aire frío asociada a dicho sistema reforzará las condiciones de ambiente frío sobre regiones de el norte |

# **Apéndice C:**
# Ejemplo de patrones

En este apartado se muestran ejemplos de patrones obtenidos en la fase de identificación de patrones del capítulo 5 de este documento.

## Patrones de la categoría de ciclones

| **Categoría** | **Patrones** |
|---|---|
| Ciclones | DA NC CN ubicar SP Z NC SP DA NC SP NP Fc NP |
| | DA NC CN ubicar SP Z NC SP DA NC SP NP Fc NP Fc SP NC AQ VM SP Z NC |
| | DA NP VM NC RG nublar SP NC Fc Z SP NC SP NC |
| | DA NP VM SP NC RG nublar SP NC Fc Z SP NC SP NC |
| | DA NC CN VM SP Z kilómetro SP DA NC SP NP |
| | NC CN VM SP Z kilómetro SP DA NC SP NP Fc NP |
| | DA NC CN VM SP Z NC SP DA NC SP NP Fc NP Fc SP viento AQ VM SP Z NC |
| | DA NC VM viento AQ VM SP Z NC Fh NC Fc NC SP SP Z NC Fh NC CC NC SP DA NC |

## Patrones de la categoría de heladas

| Categoría | Patrones |
|---|---|
| Heladas | DA NP VM NC RG nublar SP NC Fc Z SP NC SP NC |
| | DA NP VM NC RG nublar SP NC Fc Z SP NC SP NC AQ SP NP |
| | DA NP VM NC despejar DA AQ NC SP DA NC |
| | DA NP VM NC despejar SP NC VM CC Z SP NC SP NC |
| | DA NP VM NC despejar SP NC VM Fc Z SP NC SP NC AQ SP Z NC SP NP |
| | DA NP prever NC SP RG Z NC |
| | DA NP CN prever NC RG VM SP NC |
| | DA NP VM CS DA temperatura AQ SP DA AO Z NC CN VM SP NP Fc NP Fc SP Z NC |
| | DA NP VM CS SP DA AO Z NC CN VM DA temperatura AQ SP NP Fc NP Fc SP Z NC |
| | DA NP CN VM NC RG VM SP NC Fc Z SP probabilidad SP NC AQ SP NP |
| | DA NP VM NC RG VM SP NC Fc Z SP probabilidad SP NC AQ SP NP |
| | DA NP VM NC RG VM Fc Z SP NC SP NC AQ SP Z milímetro SP NP Fc NP |

## Patrones de la categoría de Inundación

| Categoría | Patrones |
|---|---|
| Inundación | DA NC VS DA AQ NC SP NP Fc PR tener DI SP DA NC AQ RG AQ SP DA NC CC VS VM SP DP AQ NC SP NC |
| | NC PR poder VM PP SP DA NC |
| | DA AO NC PR CN hacer DD NC VS SP DA NC SP Z |
| | DA NC CN haber VM Z NC SP NC SP NC |
| | DA NC VM SP devastar NC SP NC CC SP AQ NC SP NC |
| | DA NC AQ haber VM SP DA NC AQ PR RN |
| | DA NC CN haber VM Z NC SP NC SP NC |
| | DA AO AQ tormenta PR VM DA AQ NC AQ SP DI NC |
| | DA NC CC NC SP DA país VM NC AQ SP NC SP NP CC NP |
| | DA NC VM CS DA NC AQ tropical CN VM SP Z NC SP DA NC SP NP |

## Patrones de la categoría de precipitación pluvial

| Categoría | Patrones |
|---|---|
| Precipitación Pluvial | DA NP VM NC RG nublar SP NC Fc NC SP NC |
| | DA NP VM NC RG nublar SP NC Fc NC SP NC AQ SP Z SP NP |
| | DA NP VM NC RG nublar SP NC Fc NC SP NC |
| | DA NC VM registrar NC SP RG Z SP AQ NC SP NC SP NC SP NP |
| | DA NP tener NC RG VM SP NC Fc NC SP NC |
| | DA NP CN tener NC RG VM SP NC Fc Z SP NC SP NC AQ SP NP CC NP Fc NC AQ SP AQ SP DA NC CC AQ SP DA NC SP AQ NC SP DA NC |
| | DA NP prever CS DA NC NC Z CN VM SP NC SP AQ SP DA NC SP DA NC |
| | DA NC afectar SP DA NC AQ VS NP Fc NP Fc NP |
| | DA NC VM NC SP NC AQ SP NC SP SP Z NC SP hora SP DA NC NC SP NP CC NP |

## Patrones de la categoría de sequía

| Categoría | Patrones |
|---|---|
| Sequía | DA NC tener DA NC SP VM PP SP DA NC |
| | DA NC SP DA Fz NP Fz poder VM DP AO NC |
| | DA NC AQ VA VM DI NC SP Z NC SP poder VM DA NC NC SP DA NC SP NP |
| | DA NC AQ PR CN haber VM SP DA NC |
| | DA Fz NP Fz haber VM NC SP NC SP PR CN VM SP DA NP |
| | DA NC AQ haber VM DI NC SP Z NC SP VM VM DA NC NC SP DA NC SP NP |
| | DA NC SP DA sequía CC DA NC SP NC |
| | DA NC AQ VA VM DI sequía SP Z NC SP VM VM DA NC NC SP DA NC SP NP |
| | DA AQ sequía PR VM DA NC |
| | DA NC AQ SP DA agua SP NC SP DA NC |
| | DA AQ sequía PR CN VA VM SP DA NC |
| | DA AQ NC VS VM SP DA AQ NC registrar SP Z Fc DA NC NC Z SP RG NC SP DI DA NC SP DA |
| | NC SP DA NC VM SP DI NC climático AQ PR AQ NC SP DI NC |

## Patrones de la categoría de sismos

| Categoría | Patrones |
|---|---|
| Sismos | DA NC AQ tener DI NC SP Z |
| | DA NC poder VM SP DI NC |
| | DA NC AQ VM CS decir NC VM NC SP NC |
| | DA NC ocurrir SP DA Z NC |
| | DA NP RN VM DA NC SP CS ocurrir RG NC SP DA AQ NC CC NC |
| | DA NP VM CS DA NC ocurrir SP Z NC SP DA NC SP NP |
| | NC CN registrar SP DA Z NC |
| | DA NC SP NC SP DA NC NP registrar Z NC SP AQ NC |
| | DA NC AQ informar CS Fc SP DA AQ NC CN VM RG SP Z NC SP NC CC NC SP NC |
| | DA NP RN VM DA NC SP CS ocurrir RG NC SP DA AQ NC CC NC |
| | DA NP VM sismo SP NC Z CC Z |
| | DA NC SP AQ NC CC SP RG SP Z sismo Fc DA PR VM SP DA NC AQ SP VM DI NC SP NC |
| | DA NC SP DA NC AQ CN VM SP Z NC SP DA AQ ciudad SP NP Fc PR SP DA AO NC VA VM NC SP NC SP AQ NC |
| | DA NC ocurrir SP Z NC SP DA NC SP NP |

## Patrones de la categoría de tormentas

| Categoría | Patrones |
|---|---|
| Tormentas | DA NP tener NC RG VM SP NC |
| | DA NP tener NC RG VM SP NC Fc NC SP NC |
| | DA NP VM NC RG nublar SP NC Fc NC |
| | DA NP VM NC nublar Fc NC SP NC AQ SP Z SP NP |
| | DA NC SP DA AO NC haber VM NC SP DA NC SP DI NC |
| | DA NP haber NC RG VM Fc NC SP NC |
| | DA NC SP NC AQ VM SP DA NC mantener NC RG AQ SP DA NC |
| | DA NC SP DA AO NC VA generar NC SP DA NC SP DI NC |
| | DA NC SP DA AO NC haber VM NC SP DA NC SP DI NC |
| | DA NP VM NC VM Fc NC SP tormenta AQ SP Z SP NP CC NP |
| | DA NP VM NC VM Fc NC SP tormenta AQ SP Z SP NP CC NP |
| | DA NP VM CS DA NC AQ SP DA AO Z hora CN VM SP NP Fc NP |

# Patrones de la categoría de Tsunamis

| Categoría | Patrones |
|---|---|
| Tsunamis | DA NC poder VM DA NC SP NP SP VM DP NC |
| | DA NP informar CS SP DA NC |
| | DA AQ NC PR generar DI NC SP NC AQ Fc NC SP NC CC NC SP DI NC |
| | DA NC SP NP generar AQ NC SP NC |
| | DA NC RN VM DA NC SP generar DI NC SP DA NC |
| | DA NP emitir DP NC AQ SP DA NC SP Z NC SP NC |
| | DA NC CN haber VM SP DA NC SP NC |
| | DA NC CN VM Z NC SP CS DI AQ NC CC DI NC tsunami VM NC SP DA NC |
| | DA NC nuclear VM SP DA NC CC DA NC |
| | DA NC nuclear VM SP DA NC |

# Referencias


Aguilar Galicia, H., Sidorov, G., & Nikolaevna Ledeneva, Y. (2012). *Extracción automática de información semántica basada en estructuras sintácticas*. Instituto Politécnico Nacional.

Alarcos Llorach, E. (2000). Gramática de la Lengua Española. Retrieved May 18, 2015, from http://www.biblioises.com.ar/Contenido/400/460/RAE-Gramatica-de-la-lengua.pdf

Andersen, P. M., Hayes, P. J., Huettner, A. K., Schmandt, L. M., Nirenburg, I. B., & Weinstein, S. P. (1992). Extracción automática de hechos de los comunicados de prensa para generar noticias históricas. *Proceedings of the Third Conference on Applied Natural Language Processing* -, 170–177. http://doi.org/10.3115/974499.974531

Boiński, T., & Brzeski, A. (2014). Hacia la extracción de hechos de textos en el lenguaje polaco. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(8), 5231–5234.

Català Roig, N., & Castell Ariño, N. (1997). Construcción automática de diccionarios de patrones de extracción de información. *Procesamiento Del Lenguaje Natural*, *21*, 123–136.

Centro Nacional de Prevención de Desastres. (2001). *Diagnóstico de Peligros e Identificación de Riesgos de Desastres en México*.

Dali, L., & Fortuna, B. (2008). Extraccción de tripleta de oraciones usando Svm. In *SiKDD*. Ljubljana, Slovenia.: 17 octubre 2008.

Elvira, J. (n.d.). Verbos defectivos en español, *I*.

Gálvez Rojas, S., & Mora Mata, M. A. (2005). *Java a Tope: Traductores Y Compiladores Con Lex/Yacc, Jflex/Cup Y Javacc. Edición Electrónica*.

Garera, N., & Yarowsky, D. (2009). Modelos estructurales, transitivos y latentes de extracción de hechos biográficos. *Proceedings of the 12th Conference of the European Chapter of the*

*ACL*, (April), 300–308. http://doi.org/10.3115/1609067.1609100

Gelbukh, A. (2010). Procesamiento de Lenguaje Natural y sus Aplicaciones. *Enero-Junio 2010*, 6.

Hernández Jiménez, M. P., Marines Zane, M. Á., & Montes San Agustín, R. (2012). *Sistemas de Extracción de Hechos de las Personalidades Históricas en México.* Universidad Autónoma Del Estado De México.

Hernando Cuadrado, L. A. (1995). *Introducción a la teoría y estructura del lenguaje*. España: Editorial Verbum. Retrieved from http://goo.gl/sf7lce

Hofliger, R., Mahul, O., Ghesquiere, F., & Perez, S. (2012). FONDEN. El Fondo de Desastres Naturales de México-Una Reseña. Retrieved from http://www.proteccioncivil.gob.mx/work/models/ProteccionCivil/Resource/469/1/images/LibroFonden_versionEsp.pdf

Jonathan Hedley. (n.d.). jsoup: Java HTML Parser.

Kastner, I., & Monz, C. (2009). Extracción automática de hechos clave de documentos individuales en artículos de prensa. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 415–423). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1609067.1609113

López Cubino, R., López Sobrino, B., & Bernabeu Morón, N. (2009). *La noticia y el reportaje*. España: Mediascopio Serie Guías y Talleres © CIDE MEDIASCOPIO COORDINACIÓN.

Mateos, D. E., & Gonzalez, T. A. (2014). Lengua castellana y literatura. In *Mc Graw Hill* (pp. 77–94). España: S.A. MCGRAW-HILL / INTERAMERICANA DE ESPAÑA.

Messineo, C., & Klein, H. E. M. (2005). Expresión de la TRAYECTORIA en verbos de movimiento y posición en Toba ( flia guaycurú ).

Nakashole, N. T. (2013). *Extracción automática de los hechos, relaciones y entidades para la web a gran escala de la población base de conocimientos*. Universidad de Saarland. Retrieved from http://scidok.sulb.uni-saarland.de/volltexte/2013/5054/

Real Academia Española. (n.d.). Heurística. Retrieved February 25, 2015, from http://buscon.rae.es/drae/srv/search?val=heur?sticas

Sidorov, G. (n.d.). basada en los verbos de habla, *68*(Cic), 137–153.

Sidorov, G., & Herrera-de-la-cruz, J. A. (2011). Algoritmo heurístico para la extracción de

hechos de uso modelo relacional y datos sintácticos. *Springer-Verlag*, 328–337.

Talmy, L. (2000). The relation of grammar to cognition. *Toward a Cognitive Semantics - Vol. 1 Chap. 1*, *1*, 165–205. http://doi.org/10.3115/980262.980266

Tolosa, G. (2008). *Introducción a la Recuperación de Información*. Argentina.

Wróblewska, A., & Sydow, M. (2012). DEBORA: Extracción de triples entidad-relación basado en dependencias de textos polacos de dominio abierto. In 20th International Symposium & C. ISMIS 2012, Macau (Eds.), *Springer-Verlag* (pp. 155–161). Berlin Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-34624-8_19