Universidad Autónoma Metropolitana
Unidad Azcapotzalco
División de Ciencias Básicas e Ingeniería
Licenciatura en Ingeniería en Computación

Reporte de Proyecto Tecnológico

Extracción automática de eventos indicadores a partir de noticias en español

Presenta:
Ariatna Quinto Martínez
2113034676

Asesores

Dr. José Alejandro Reyes Ortiz
Doctor en Ciencias de la Computación
Departamento de Sistemas

Dra. Angeles Belém Priego Sánchez
Doctora en Ciencias del Lenguaje
Departamento de Sistemas

Trimestre 2017- Otoño

11 de diciembre del 2017

Declaratoria

Yo, José Alejandro Reyes Ortiz, declaro que aprobé el contenido del presente Reporte de Proyecto de Integración y doy mi autorización para su publicación en la Biblioteca Digital, así como en el Repositorio Institucional de UAM Azcapotzalco.

Dr. José Alejandro Reyes Ortiz

Yo, Angeles Belém Priego Sánchez, declaro que aprobé el contenido del presente Reporte de Proyecto de Integración y doy mi autorización para su publicación en la Biblioteca Digital, así como en el Repositorio Institucional de UAM Azcapotzalco.

Dra. Angeles Belém Priego Sánchez

Yo, Ariatna Quinto Martínez, doy mi autorización a la Coordinación de Servicios de Información de la Universidad Autónoma Metropolitana, Unidad Azcapotzalco, para publicar el presente documento en la Biblioteca Digital, así como en el Repositorio Institucional de UAM Azcapotzalco.

Ariatna Quinto Martínez

Resumen

Actualmente, se ha notado el incremento exorbitante de la información electrónica, claro ejemplo es la Web, la cual se ha convertido de fácil acceso. Ésta es posible estudiarla para saber los fenómenos que suceden a nivel de la lengua y a partir de ella se puede extraer meta información. Por tal motivo, en este proyecto se han seleccionado artículos periodísticos en formato electrónico, del gran repositorio de información que es la Web, se ha filtrado una parte de la información existente y posteriormente se tendrán únicamente estadísticas que ocurren alrededor del mundo. Una vez que se han extraído estas estadísticas, los eventos indicadores a través de patrones lingüísticos, los resultados se podrán visualizar mediante un sistema web que mantendrá informados a los usuarios.

En el Diccionario de la Real Academia [1], una de las acepciones para "evento" dice que es un suceso importante. De esta manera, este proyecto retoma esa idea para mostrar un suceso importante que surge alrededor del mundo a través de los relatos periodísticos que se encuentran en la web. Sin embargo, se sabe que dar una definición concisa de evento es difícil. Por lo tanto, en este proyecto se trata a un suceso como cualquier tipo de situación o acontecimiento que ocurre, restringiendo a eventos relacionados con la proporcionalidad de uno con respecto a un total, es decir porcentajes relacionados a un suceso. En los ejemplos (1) y (2) se pueden observar enunciados que muestran eventos indicadores.

(1) *En México, cerca del 88% de la energía primaria que se consume proviene del petróleo.*

(2) *La mitad de la población mundial está concentrada en tan solo seis países.*

En (1) el evento indicador de porcentaje está explícitamente determinado por su signo ortográfico. Mientras que en (2) se nota un evento indicador, *la mitad de la población mundial*, pero no está explícitamente denotado por un signo ortográfico. Por lo que la tarea de extracción de eventos indicadores se convierte más complicada a medida que el lenguaje se desarrolla con diferentes factores de pensamiento, es decir, a medida que el lenguaje es abundante, enriquecido y creativo, la tarea se complica.

Típicamente, los eventos son expresados por verbos conjugados o en infinitivo, predicados en general y frases preposicionales. Sin embargo, como se notará, en este proyecto no se cumple con esos acuerdos, ya que se pueden encontrar eventos expresados mediante un signo de puntuación (%), partes que dividen un todo (mitad, tercio, etc.), entre otras expresiones. Sobre un texto plano, se pretende identificar eventos indicadores mediante diferentes patrones lingüísticos que serán identificados y son los que ayudaran a la extracción de diferentes sucesos acontecidos alrededor del mundo. Se puede delimitar entre las etiquetas <indicador> y </indicador>, el inicio y fin del indicador porcentual con el fin de poder extraer toda la idea que contiene al evento, en (3) se puede observar un ejemplo. Esto mediante un sistema que segmente a los párrafos en oraciones.

(3) *La <indicador>mitad</indicador> de la población mundial está concentrada en tan solo seis países.*

La demanda creciente de información sobre diversos aspectos de la realidad en el mundo y nuestra sociedad, fue una de las razones que impulso el desarrollo de este proyecto. Debido a que el público, en general, consulta estadísticas para diversos fines, destacándose el de conocer aspectos esenciales de la realidad nacional e internacional, como parte de la cultura general del ciudadano en el mundo actual.

En el contexto de este proyecto se contará con un esquema que establece reglas claras con guías de cómo se deben identificar los eventos indicadores a partir de diferentes patrones lingüísticos, con el objetivo de reducir las ambigüedades al mínimo.

En este reporte se detallará la metodología utilizada para el desarrollo del sistema web, el cual mostrará las diferentes oraciones extraídas del corpus, periodístico en español, a partir de los diferentes patrones lingüísticos extraídos. La metodología utilizada, a grandes rasgos, para llevar a cabo el diseño del sistema web fue la siguiente:

Desarrollo de tres módulos principales, procesamiento, extracción de eventos indicadores y visualización de dichos eventos. La etapa de procesamiento permite, de cierto modo, limpiar los textos del corpus. La extracción de eventos indicadores se lleva a cabo mediante el diseño de patrones lingüísticos que cubren los tres niveles de la lengua, morfológico, sintáctico y semántico, uno de los principales objetivos de este trabajo. A lo largo de este documento se presentarán las etapas que intervienen en este proceso de extracción.

Una vez obtenidos los patrones lingüísticos, a visualizar, se diseña e implementa un sistema web que los muestra. Proporcionándole al usuario una interacción con éstos, la interacción se realiza de acuerdo a los diferentes patrones lingüísticos extraídos, es decir, los resultados están presentados conforme a cada patrón extraído.

Tabla de contenido

CAPÍTULO 1. Acercamiento teórico 1
1.1 Introducción 1
1.2 Antecedentes 2
1.3 Justificación 4
1.4 Objetivos 5
CAPÍTULO 2. Marco teórico 6
2.3 NetBeans IDE 8.1 6
2.4 Java 6
2.5 JavaScript 7
2.6 ¿Qué es HTML? 7
2.7 ¿Qué es CSS? 10
CAPÍTULO 3. Desarrollo del proyecto 11
3.1 Módulo de procesamiento 11
3.2 Módulo de extracción de eventos 13
3.3 Módulo de visualización 18
CAPÍTULO 4. Resultados 24
4.1 Resultados 24
4.2 Análisis y Resultados 54
Conclusiones 55
Referencias bibliográficas 56
Apéndices 57
Apéndice A. Publicación de Articulo en Congreso 57
Apéndice B. Palabras cerradas en Español 66

Apéndice C. Clase para leer archivo y obtener el resultado a visualizar..... 72
Apéndice D. Código jsp para visualización de sistema web 72

Índice de Figuras

Figura 1. Esquema HTML y CSS. Imagen tomada de la referencia [16] 7
Figura 2. Opción de Inicio 18
Figura 3. Opción ACERCA DE del sistema web. 19
Figura 4. Opción ELABORADO POR del sistema web 20
Figura 5. Opción CONTACTOS del sistema web 21
Figura 6. Menú para elegir opciones de eventos indicadores en el sistema web 22
Figura 7. Código para el despliegue de resultados 23
Figura 8. Prueba uno del evento indicador % en el sistema web 24
Figura 9. Prueba dos del evento indicador % en el sistema web..... 25
Figura 10. Prueba tres del evento indicador % en el sistema web .. 26
Figura 11. Prueba cuatro del evento indicador % en el sistema web 27
Figura 12. Prueba cinco del evento indicador % en el sistema web 28
Figura 13. Prueba uno del evento indicador por ciento en el sistema web 29
Figura 14. Prueba dos del evento indicador por ciento en el sistema web 30
Figura 15. Prueba tres del evento indicador por ciento en el sistema web 31
Figura 16. Prueba cuatro del evento indicador por ciento en el sistema web 32

Figura 17. Prueba cinco del evento indicador por ciento en el sistema web 33
Figura 18. Prueba uno del evento indicador la mitad en el sistema web 34
Figura 19. Prueba dos del evento indicador la mitad en el sistema web 35
Figura 20. Prueba tres del evento indicador la mitad en el sistema web 36
Figura 21. Prueba cuatro del evento indicador la mitad en el sistema web 37
Figura 22. Prueba cinco del evento indicador la mitad en el sistema web 38
Figura 23. Prueba uno del evento indicador un tercio en el sistema web 39
Figura 24. Prueba dos del evento indicador un tercio en el sistema web 40
Figura 25. Prueba tres del evento indicador un tercio en el sistema web 41
Figura 26. Prueba cuatro del evento indicador un tercio en el sistema web 42
Figura 27. Prueba cinco del evento indicador un tercio en el sistema web 43
Figura 28. Prueba uno del evento indicador una cuarta parte en el sistema web 44
Figura 29. Prueba dos del evento indicador una cuarta parte en el sistema web 45
Figura 30. Prueba tres del evento indicador una cuarta parte en el sistema web 46
Figura 31. Prueba cuatro del evento indicador una cuarta parte en el sistema web 47
Figura 32. Prueba cinco del evento indicador una cuarta parte en el sistema web 48

Figura 33. Prueba uno del evento indicador un total de en el sistema web 49
Figura 34. Prueba dos del evento indicador un total de en el sistema web 50
Figura 35. Prueba tres del evento indicador un total de en el sistema web 51
Figura 36. Prueba cuatro del evento indicador un total de en el sistema web 52
Figura 37. Prueba cinco del evento indicador un total de en el sistema web 53

Índice de Tablas

Tabla 1. Etiquetas más comunes de HTML..... 8
Tabla 2. Descripción del corpus periodístico utilizado. 12
Tabla 3. Descripción de las oraciones obtenidas. 13
Tabla 4. Patrones lingüísticos usando explícitamente el porcentaje. 14
Tabla 5. Patrones lingüísticos relacionados implícitamente con un porcentaje. 15
Tabla 6. Resultados extraídos (eventos indicadores)..... 17
Tabla 7. Estadísticas de resultados extraídos (eventos indicadores) 17

CAPÍTULO 1. Acercamiento teórico

1.1 Introducción

El Procesamiento de Lenguaje Natural (denotado por PLN), es una disciplina de la Inteligencia Artificial que trata la formulación e investigación de mecanismos de computación para la comunicación entre personas y máquinas, mediante el uso de Lenguajes Naturales. Dichos lenguajes son utilizados para la comunicación ya sea de forma escrita, hablada o en forma de signos [2]. Entre las tareas que realiza el PLN encontramos la extracción automática de eventos, cuyo objetivo es capturar ciertas partes relevantes de un texto.

En el análisis del lenguaje se estudia la estructura del lenguaje a cuatro niveles [2]:

• Análisis morfológico: El análisis de las palabras para extraer raíces, rasgos flexivos, unidades léxicas compuestas y otros fenómenos.

• Análisis sintáctico. El análisis de la estructura sintáctica de la frase mediante una gramática de la lengua en cuestión.

• Análisis semántico. La extracción del significado (o posibles significados) de la frase.

• Análisis pragmático. El análisis de los significados más allá de los límites de la frase, por ejemplo, para determinar los antecedentes referenciales de los pronombres.

En este proyecto se aborda la tarea de la extracción automática de eventos a partir de noticias en español. En particular, los eventos son los indicadores que representan la proporcionalidad de un evento con respecto a un total que aparecen en una colección de información (noticias), es decir, éstos son los porcentajes relacionados con algún suceso. Para el desarrollo del proyecto se hace uso únicamente de tres niveles del lenguaje, el morfológico, el sintáctico y el semántico. Además, se selecciona el género periodístico debido a que es un tipo de escritura estándar y homogénea, en otras palabras, cualquier hablante nativo que lea el contenido de una nota periodística, lo entiende. Inclusive, la mayoría de las personas tiene acceso a un periódico, en formato papel o digital.

1.2 Antecedentes

A continuación, se presentan algunos trabajos relacionados con el tema de extracción automática de eventos en textos, algunos en particular como son los textos periodísticos.

De igual forma se tomó en cuenta el desarrollo de algunos softwares realizados específicamente con la extracción de terminología de textos.

1.2.1 Tesis

1. Sistema Web para identificar eventos y actores de textos periodísticos [3].

En dicho proyecto se diseñó e implemento un Sistema Web para la anotación semántica de actores y eventos a partir de un corpus de textos periodísticos mexicanos aplicando técnicas de minería de textos haciendo uso de características sintácticas, semánticas y contextuales.

La diferencia del proyecto [3] con respecto a la propuesta que se busca desarrollar es que no se implementará minería de datos, de igual forma no se extraen actores. Una similitud importante es que la extracción se hará de textos periodísticos mexicanos.

2. Reconocimiento automático de eventos en textos [4].

En el contexto de este trabajo se cuenta con un esquema que establece reglas claras con guías de cómo se deben marcar los eventos, a manera de reducir las ambigüedades al mínimo.

En el proyecto [4] se propone identificar los eventos o sucesos mencionados en textos y determinar el momento en que estos ocurrieron, si es que efectivamente ocurrieron. Lo cual, en el proyecto propuesto, difiere ya que solo se realizará la extracción de textos periodísticos y las estadísticas que se visualizaran están, de cierto modo, comprobadas ya que está de por medio una persona especialista en el tema, por lo tanto, no hay ambigüedad en las estadísticas.

1.2.2 Artículos

3. Hacia una identificación de la similitud verbal para la extracción de eventos. [5].

La función de este proyecto es extraer eventos mediante la interpretación de las entidades y relaciones que un texto posee mediante lo aplicación de *kybots*, heurísticas que explotan documentos con información lingüística, esta información representa un patrón de extracción de información relevante para un cierto evento.

La diferencia sustancial entre lo planteado en la propuesta y la descrita en [5], es la utilización de verbos para extraer los eventos relacionados a éstos y la forma de visualización, ya que los resultados quedan a nivel texto. En el caso, del proyecto planteado se buscarán patrones que abarquen más allá de los verbos y los resultados serán mostrados en un sistema web.

4. Extracción automática de metadatos a partir de objetos de aprendizaje en un repositorio institucional [6].

En el artículo [6] se busca propiciar el uso de los repositorios institucionales, a través de la búsqueda y consulta de material educativo. Para esto, es importante contar con una buena descripción de los Objetos de Aprendizaje que conforman el repositorio, a partir de la calidad de los metadatos descriptivos, permite la recuperación de aquellos objetos que mejor satisfagan las necesidades de información del usuario, teniendo en cuenta sus características y preferencias individuales.

Una de las principales diferencias con el proyecto a desarrollar es que en el artículo [6] la extracción automática es de metadatos, sin embargo, igual que lo que se plantea es la búsqueda de ciertas características relevantes para darle información importante y verídica al usuario final.

1.2.3 Software

5. JASPER (*Journalist's Assistant for Preparing Earnings Reports*) [7].

El software JASPER, es un sistema para la extracción de ciertas piezas clave de información de un rango limitado de texto, éste es presentado en [7]. Este sistema está basado en el uso de plantillas, técnicas de comprensión parcial y procedimientos heurísticos para extracción de información. Esta información, puede ser utilizada de varias maneras, como rellenar valores en una base de datos, generar resúmenes del texto de entrada, entre otras.

La similitud del sistema descrito en [7] con el que se desarrollará, es justamente la extracción de piezas claves de algún corpus, en este proyecto denominadas como eventos. Los eventos, utilizados en [7], son los comunicados de prensa para generar historias de noticias. Con respecto al proyecto a implementar, los eventos serán basados en indicadores (porcentajes) que aparecen en una nota periodística para generar un sistema que las muestre.

6. TES (*Terminology Extraction Suite*) [8].

El software [8] es una herramienta desarrollada para la extracción automática de terminología, que permite extraer términos y buscar automáticamente equivalentes de traducción. La herramienta está escrita en Perl, con interfaces gráficas implementadas en Tk.

Una de las similitudes con el proyecto a desarrollar, se basa en que se realizará la extracción automática de eventos indicadores. Estos eventos se pueden ver como términos con

características diferentes y son obtenidos a partir de patrones lingüísticos, de igual manera que en [8], para la extracción de términos se realiza mediante patrones lingüísticos.

1.3 Justificación

Debido a la basta cantidad de información que actualmente encontramos en la Web, ésta la podemos utilizar y procesar de modo que se pueda emplear para ciertas tareas del PLN, una de ellas es la extracción de información relevante de un texto. Además de que es posible extraer conocimiento de toda esta información. Uno de los medios de comunicación que proporciona información es el periódico, que especialmente en los últimos años su acceso ha sido en formato digital. Esto debido a los avances tecnológicos, como es internet. Razón por la cual, en este proyecto se ha decidido trabajar con este género textual. A través de los periódicos se ha podido llegar incluso a más gente y mantener un ritmo de actualización de los datos mucho más intenso que antes, siendo hoy imposible esperar de un día para otro para conocer noticias.

Lo interesante de los periódicos es que cuando hablamos de una sociedad más o menos compleja, podemos encontrar distintos tipos de periódicos que dan con el perfil de grupos sociales particulares, de grupos de edad, de regiones geográficas, de actividades laborales, de intereses específicos como deportes, internacionales, espectáculos o política. Debido a que la sociedad cada vez vive de una manera más acelerada, se ha decidido extraer información de las noticias y a través de esta extracción dar a conocer datos estadísticos que la búsqueda de eventos indicadores proporcionan como resultado. Lo que sirve para informar de manera más concisa y directa cierta información, ésta representa la proporcionalidad de un evento, sin la necesidad de que las personas lean toda la nota periodística, o más notas, para conocer datos importantes y relevantes. El sistema web, en el que se muestran los resultados de la extracción de eventos indicadores, sirve como medio de información resumida, lo cual beneficia a la sociedad que desea estar informada.

1.4 Objetivos

1.4.1 Objetivo General

Diseñar un sistema web para la extracción y visualización de eventos indicadores a partir de un corpus de notas periodísticas en español utilizando patrones lingüísticos.

1.4.2 Objetivos Específicos

- Diseñar e implementar un módulo de procesamiento de notas periodísticas, el cual incluye la limpieza del corpus, el análisis morfológico (etiquetado) y sintáctico (segmentación en oraciones).
- Diseñar e implementar patrones lingüísticos para la extracción de eventos indicadores que representan la proporcionalidad de un evento con respecto a un total.
- Diseñar e implementar un sistema web que permita visualizar e integrar los resultados obtenidos en proceso de la extracción de eventos indicadores.

CAPÍTULO 2. Marco teórico

En este capítulo se proporciona conocimiento para desarrollar un sistema web y se menciona como es el uso de algunas herramientas necesarias para llevar a cabo la implementación de este sistema.

2.3 NetBeans IDE 8.1

2.3.1 ¿Qué es NetBeans IDE 8.1?

Es una herramienta para que los programadores puedan escribir, compilar, depurar y ejecutar programas. Está escrito en Java, pero puede servir para cualquier otro lenguaje de programación. Existe además un número importante de módulos para extender el NetBeans IDE[1]. NetBeans IDE es un producto libre y gratuito sin restricciones de uso [12].

2.4 Java

2.4.1 ¿Qué es Java?

Java es un lenguaje de programación y una plataforma informática comercializada por primera vez en 1995 por *Sun Microsystems* [13].

2.4.2 Razones para utilizar Java

- Para crear aplicaciones Android.

Si se tiene el conocimiento necesario se pueden desarrollar aplicaciones para que miles de usuarios puedan descargarlas y usarlas en sus dispositivos móviles.

- Java es multiplataforma.

Se puede desarrollar una sola aplicación que funcione en cualquier plataforma, ya sea Windows, Mac o Linux sin la necesidad de pagar ninguna licencia ya que es completamente gratuito usar esta tecnología.

- Gran soporte y documentación.

La comunidad de Java tiene disponible un gran soporte y documentación para seguir aprendiendo, resolver dudas y problemas que surjan al momento de desarrollar aplicaciones.

- Java es código abierto.

[1] Un IDE es un entorno de programación que ha sido empaquetado como un programa de aplicación, o sea, consiste en un editor de código, un compilador, un depurador y un constructor de interfaz gráfica.

Los usuarios pueden estudiar, modificar y mejorar su diseño mediante la disponibilidad de su código fuente.

2.5 JavaScript

2.5.1 ¿Qué es JavaScript (JS)?

Es un lenguaje ligero e interpretado, orientado a objetos con funciones de primera clase, mejor conocido como el lenguaje de script para páginas web, pero también usado en muchos entornos sin navegador. Es un lenguaje script multi-paradigma, basado en prototipos, dinámico, soporta estilos de programación funcional, orientada a objetos e imperativa [14].

En la Figura 1, se presentan de manera gráfica dos importantes herramientas que se detallarán, de manera general, en este capítulo ya que estas proporcionan recursos para darle un mejor diseño y vista al sistema web.

Figura 1. Esquema HTML y CSS.

2.6 ¿Qué es HTML?

Es el acrónimo en inglés de *HyperText Markup Language* y es un lenguaje de marcas, es decir, consta de texto, que define los contenidos reales de la página web, y de marcas especiales (también conocidas como etiquetas o *tags*) que permiten dar "significado" al texto o contenido, así como indicar algún tratamiento especial sobre dicho texto [15].

En la Tabla 1**,** se muestran algunas etiquetas comunes para el uso correcto de HTML, ya que se debe tener especial cuidado en el uso de dichas etiquetas en esta tabla se agrupan de acuerdo a la función que realizan.

Tabla 1. Etiquetas más comunes de HTML

Concepto	Etiqueta	Descripción	Atributos principales
Elemento raíz	html	engloba todo el documento	Lang
Metadatos	head	delimita el encabezado del documento	
	title	título del documento (se muestra en la pestaña del navegador)	
	link /	enlace a otros archivos (hoja de estilo, etc.)	href, rel, media, type, title
	meta /	metainformación sobre el documento	name, content, charset
	style	hoja de estilo incluida en el documento	type, title
Secciones	body	delimita el cuerpo del documento	
	article	Artículo	
	section	Sección	
	nav	Navegación	
	aside	Lateral	
	h1 a h6	encabezado (de nivel 1 a 6)	
	header	Cabecera	
	footer	Pie	
Contenido (bloques)	p	Párrafo	
	hr /	Separador	
	div	división	
	pre	texto preformateado	
	main	principal	
	figure	ilustración	
	figcaption	pie de ilustración	
Texto (en línea)	br /	salto de línea	
	a	hiperenlace	href, target, download, rel, type
	br /	salto de línea	
	sub	subíndice	
	sup	superíndice	
	data	datos	Value
	time	fecha y hora	Datetime

	var	variable (de programa de ordenador)	
Contenido incrustado	img /	imagen	alt, src, usemap, ismap, width, height
	iframe	marco incrustado en el documento	src, srcdoc, name, sandbox, width, height
	video		src, poster, preload, autoplay, loop, muted, controls, width, height
	audio		src, preload, autoplay, loop, muted, controls
	map	mapa de imagen	Name
	area /	área en mapa de imagen	alt, coords, href, hreflang, rel, shape, target, type
Listas	ol	lista ordenada	reversed, start, type
	ul	lista no ordenada	
	li	elemento de lista (ordenada o no ordenada)	Value
Tablas	table	tabla	Border
	caption	leyenda de tabla	
	thead	cabecera de tabla (grupo de filas)	
	tr	fila	
	td	celda	colspan, rowspan, headers
	col	columna	Span
Formularios	form	formulario	accept-charset, action, autocomplete, enctype, method, target
	label	etiqueta de un control	form, for
	input /	control (hay varios tipos)	type (submit, reset, button, text, password, number, search, tel, url, email, date, time, color, range, file, image, hidden, etc), name, value, checked, selected, width,

			height, size, maxlength, ...
	button	botón	name, type, value, form
	select	caja de lista	name, multiple, size, ...
	datalist		
	option	opción de caja de lista	label, selected, value
	textarea	área de texto	name, cols, rows, ...
Scripts	script	script	src, type, charset, async, defer
	noscript	contenido a mostrar en navegadores que no admiten <script>	
	template	plantillas utilizables por scripts	
	canvas	zona de dibujo utilizable por script	width, height
Otros	!DOCTYPE	tipo de documento (versión de html empleada)	
	<!-- ... -->	comentario (sólo visible en el código fuente)	

2.7 ¿Qué es CSS?

Las CSS (*coding style sheets*) u hojas de estilo en cascada son los archivos responsables de definir la **apariencia** de una página web. Facilitan la gestión y la apariencia corporativa de un sitio web, ya que todas las páginas de un sitio pueden compartir la misma hoja de estilos y los cambios de apariencia quedan centralizados en estos archivos. Además, separan la presentación del contenido por lo que podemos cambiar la estética de una página sin tener que cambiar su contenido (HTML). Una hoja de estilos no es más que un archivo de texto con extensión .css. [15].

Una hoja de estilos define una o más reglas que se aplicaran a los elementos que cumplen dicha regla. Cada regla se compone de dos partes:

- Selector: indica a que elementos se va a aplicar la regla. Pueden escribirse varios selectores para la misma regla separados por comas (,).
- Lista de declaraciones: los estilos que se van a aplicar a los elementos que cumplen la regla. Son pares **propiedad: valor,** separados por punto y coma **(;)**

CAPÍTULO 3. Desarrollo del proyecto

El objetivo de este capítulo es dar a conocer la metodología desarrollada para llegar a la extracción automática de eventos indicadores. Dicho capítulo, está dividido en tres secciones, correspondientes a los módulos del sistema, cada una con diferentes subsecciones. La primera subsección, el de procesamiento, básicamente se realizan dos actividades importantes como es el preprocesado del corpus, y el análisis sintáctico. En la segunda subsección, el módulo de extracción de eventos, busca obtener patrones lingüísticos e implementarlos. Por último, en la tercera subsección, se presentan los pasos para realizar el módulo de visualización, el cual tiene como objetivo la elaboración del sistema web. En nuestro caso, éste muestra los patrones obtenidos en el segundo módulo.

3.1 Módulo de procesamiento

En este módulo, como ya se mencionó, se compone de dos etapas. En cada una de ellas, se presenta una descripción y resultados obtenidos al procesar el corpus.

3.1.1 Limpieza del Corpus.

En este proceso se eliminó el guion medio (-) y la etiqueta "|||" que es la que separa los metadatos del corpus. Dichos metadatos, que componen a cada una de las notas periodísticas, son el título de la noticia, el lugar, la fecha y como tal la noticia. Estos datos están conformados por el siguiente formato:

Título ||| Lugar ||| Fecha ||| Noticia

El formato del corpus es presentado en el ejemplo (4)

(4) *Entregaron la Seduzac ||| Zacatecas ||| 29 de mayo de 2013 ||| Haydee Santillán Zacatecas, Zacatecas. - Se reactivan las actividades en la Secretaría de Educación en el Estado (SEDUZAC), esto después de tres semanas de que el Movimiento Magisterial Democrático efectuara un plantón a las afueras de dicha institución, en señal de protesta ante las afectaciones causadas por el gobierno tras manifestar su desacuerdo con la Reforma Educativa…*

Estos elementos, "-" y "|||", fueron eliminados debido a que se consideran irrelevantes para la extracción de los eventos indicadores y podrían ocasionar conflicto al momento de recuperar la información que se está extrayendo.

Para el caso, particular, de la eliminación de la etiqueta "|||", más que eliminación fue una sustitución por la etiqueta "TDIVT". Esto con el fin de adaptar el corpus al formato que se utilizará posteriormente. Para llevar a cabo este proceso, se utilizó el siguiente comando ejecutado en Linux:

```
awk[2] '{gsub(/\|\|\|/, "TDIVT", $0); print $0;}'corpusEntrada.txt> corpusSalida.txt
```

donde:

awk '{gsub(/\|\|\|/, "TDIVT", $0); print $0;}' pertenece a la instrucción del lenguaje de programación awk, que permite reemplazar la etiqueta "|||"

corpusEntrada.txt corresponde al archivo en donde se desean reemplazar las etiquetas.

corpusSalida.txt corresponde al archivo en donde se guardará el corpus preprocesado.

El corpus que se utilizó para dichos objetivos consta de 2,086,761 noticias, con un total de 24,703,655 oraciones, ver Tabla 2 que describe las características del corpus periodístico empleado.

Tabla 2. Descripción del corpus periodístico utilizado.

Característica	**Cantidad**
Número de noticas	2,086,761
Número de palabras	782,102,723

[2] AWK: es un lenguaje de programación diseñado para procesar datos basados en texto, ya sean ficheros o flujos de datos.

3.1.2 Análisis Sintáctico.

En esta tarea se dividió a los párrafos del corpus en oraciones, debido a que así fue más fácil su manipulación y las oraciones tienen una longitud más regular.

En la Tabla 3, se muestra detalladamente la descripción de las oraciones obtenidas como es el mínimo y máximo de palabras que contienen las oraciones del corpus.

Tabla 3. Descripción de las oraciones obtenidas.

Características	Cantidad
Número de oraciones	24,703,655
Número de palabras	749,805,561
Tamaño del vocabulario	2,562,706
Número mínimo de palabras en la oración	3
Número máximo de palabras en la oración	319
Promedio de palabras en la oración	33.9795841

3.2 Módulo de extracción de eventos

Este módulo se divide en dos partes esenciales, diseño e implementación de patrones lingüísticos, las cuales se describen, a continuación, a detalle.

3.2.1 Diseño de patrones lingüísticos.

En esta esta etapa se diseñarán los patrones lingüísticos que, en una etapa posterior, se implementarán con el fin de extraer los eventos indicadores. Este diseño de patrones permitirá descubrir los elementos lingüísticos empleados con frecuencia en las notas periodísticas, para ello se utilizará uno o varios modelos que sirvan como muestra para identificar y agrupar los eventos indicadores.

En la Tabla 4, se muestran ejemplos de los patrones lingüísticos más comunes encontrados mediante el proceso de generalización basado en un conjunto inicial moderadamente pequeño de experimentos.

Tabla 4. Patrones lingüísticos usando explícitamente el porcentaje.

Patrón lingüístico	**Semántica**
… NUM por ciento …	Indicador Puntual
… alrededor del NUM por ciento…	Indicador aproximado
… mayor que NUM por ciento …	Indicador de punto base
… entre NUM y NUM por ciento …	Intervalo
… de NUM A NUM por ciento …	Incremento
… hasta un NUM por ciento …	Máximo temporal
… NUM por ciento más que …	Indicador comparativo incremental
… un incremento de NUM por ciento …	Indicador comparativo incremental
… NUM por ciento menos que …	Indicador comparativo decremental
… un decremento de NUM por ciento …	Indicador comparativo decremental

Es importante aclarar que la etiqueta **NUM** se refiere a la especificación de un número en el texto periodístico en cualquiera de sus expresiones. Ejemplos de **NUM** serían los siguientes: 90, 28.3, noventa y tres, 80,3 etc.

Por otro lado, en la Tabla 4, se ha usado el texto "*por ciento*", el cual puede ser encontrado también sustituido por el símbolo ortográfico "*%*", por lo que el lector debe considerar que los patrones anteriormente mencionados pueden ocurrir con cualquiera de estas dos expresiones textuales.

Algunos ejemplos reales extraídos del corpus utilizado, son presentados en (5, 6, 7, 8, 9 y 10).

(5) *La entidad prevé ahora que el Producto Interno Bruto (PIB) mundial crezca este año un **3,1%**, en lugar del **3,4%** que proyectaba en noviembre pasado.*

(6) *Bolsa Mexicana cierra con pérdida de **0.31%.***

(7) *Poder adquisitivo del salario mínimo ha caído **9.3%.***

(8) *Incumple el **50%** de los Patrones con el Pago de las Utilidades.*

(9) *Los bonos referenciales a 10 años subían 6/32 en precio, para rendir un **2,15 por ciento.***

(10) *En tanto, la probabilidad de lluvia alcanzará **80 por ciento.***

En la Tabla 5 se muestran otros patrones morfosintácticos que expresan el uso de números fraccionarios porcentuales, pero expresados en lenguaje natural.

Tabla 5. Patrones lingüísticos relacionados implícitamente con un porcentaje.

Patrón lingüístico	**Semántica**
… un total de NUM de las/los NUM …	Indicador porcentual basado en cociente
… la mitad …	Indicador 50%
… un tercio …	Indicador 33%
… una cuarta parte …	Indicador 25%

Dos oraciones que contienen algunos de los patrones presentados en la Tabla 5, se observan en los ejemplos (11, 12).

(11) ***Una cuarta** parte de hogares poblanos apenas tiene acceso a Internet.*

(12) *Sólo **un tercio -31.8 por ciento-** de los poblanos consideró al gobierno municipal.*

3.2.2 Implementación de patrones lingüísticos.

Una vez que se han identificado los patrones lingüísticos, la implementación de éstos se realizará mediante su búsqueda en el corpus, lo que permitirá extraer la información relevante relacionada a los eventos indicadores.

Para implementar los patrones lingüísticos, descritos en el subcapítulo 3.2.1, se utilizó el comando *grep* en Linux. Este comando toma una expresión regular de la línea de comandos, lee la entrada estándar o una lista de archivos, e imprime las líneas que contengan coincidencias para la expresión regular.

Dicho comando se utilizó para este proyecto de la siguiente manera.

- `grep "%" archivodeentrada.txt > archivodesalida.txt`
- `grep "por ciento" archivodeentrada.txt > archivodesalida.txt`
- `grep "la mitad" archivodeentrada.txt > archivodesalida.txt`
- `grep "un tercio" archivodeentrada.txt > archivodesalida.txt`
- `grep "una cuarta parte" archivodeentrada.txt > archivodesalida.txt`
- `grep "un total de" archivodeentrada.txt > archivodesalida.txt`

donde:

`grep " "` pertenece a la instrucción en Linux que permite extraer la expresión regular que se encuentra entre comillas

`archivodentrada.txt` corresponde al archivo en donde se desean buscar dicha expresión regular

`archivodesalida.txt` corresponde al archivo en donde se guardarán las líneas que contengan coincidencias para la expresión regular.

Una vez que fueron extraídos del corpus los patrones lingüísticos, en la Tabla 6 se muestran numéricamente los resultados de cada patrón lingüístico diseñado.

Tabla 6. Resultados extraídos (eventos indicadores).

Patrón lingüístico	Evento indicador a buscar	Resultados extraídos	Ejemplos de eventos indicadores obtenidos
… NUM por ciento …	%	114,313	Los funcionarios de seguridad iraquí nos han engañado con sus declaraciones de que la situación ha mejorado un 80%.
	por ciento	377,901	El congresista dijo que el criterio general es que se ajuste cualquier incremento al 4 por ciento.
… la mitad …	la mitad	24,825	A diario hacen 50 panes y un tanto más de empanadas, y en ocasiones venden únicamente la mitad.
… un tercio …	un tercio	4,041	Hasta ahora la violencia no ha atemorizado a los turistas, que representan un tercio de la economía, agregó.
… una cuarta parte …	una cuarta parte	952	Aquí ya es bien conocido cuánto te dan por las prendas, siempre es una cuarta parte de lo que vale
… un total de …	un total de	59,198	En el avance del proceso de evaluación, las empresas ya validadas suman un total de 4 mil 372.

En la Tabla 7, se observan las estadísticas de palabras totales en las oraciones resultantes de los diferentes patrones lingüísticos extraídos del corpus de notas periodísticas.

Tabla 7. Estadísticas de resultados extraídos (eventos indicadores).

Patrón lingüístico	Evento indicador a buscar	Palabras totales en resultados obtenidos	Promedio de palabras que aparecen en los resultados
… NUM por ciento …	%	4,963,552	43.4199
	por ciento	17,250,476	45.6481
… la mitad …	la mitad	1,128,273	45.4490
… un tercio …	un tercio	164,431	40.6906
… una cuarta parte …	una cuarta parte	43,583	45.7804
… un total de …	un total de	2,781,328	46.9834
PROMEDIO TOTAL			44.6619

3.3 Módulo de visualización

Este módulo mostrará cómo fue diseñado el sistema web en cuanto a la vista del sistema web, así como su funcionamiento en general.

3.3.1 Vista general del sistema web

A continuación, se muestran las partes del sistema web. En las Figuras 2, 3, 4 y 5 se muestra el primer menú el cual tiene el objetivo brindar información general del sistema, así como información de contacto.

En la Figura 2 se muestra la opción de Inicio, esta opción permite hacer un refresh o dicho de otra forma recargar la página ya sea para quitar las opciones del menú principal o para visualizar nuevos resultados del menú de botones que se muestra a la izquierda del sistema web.

Figura 2. Opción de Inicio.

En la Figura 3 se puede visualizar de forma general cual es el objetivo del sistema web.

Figura 3. Opción ACERCA DE del sistema web.

En la Figura 4, se muestra la vista creada para dar conocer los nombres tanto de quien desarrollo este proyecto de integración por parte de la Universidad Autónoma Metropolita, como de los asesores de este.
En esta sección solo se proporcionan los nombres debido a que en la opción de CONTACTOS se muestra la forma de comunicación con ellos.

Figura 4. Opción ELABORADO POR del sistema web.

La vista que se muestra en la Figura 5, se desarrolló para dar a conocer al usuario la forma de contactarse con la persona encargada de este proyecto, así como con los asesores de este, también se muestran dos iconos uno de Facebook y el otro de Twitter los cuales permiten a los usuario ser enlazados al Área de Investigación en Sistemas de Información de la Universidad Autónoma Metropolitana Unidad Azcapotzalco, se consideró que estas opciones eran importantes debido a que actualmente las redes sociales tienen un papel protagónico, ya que no sólo se utilizan para la comunicación instantánea, , compartir e intercambiar información en diferentes medios, sino que también están siendo utilizadas por grandes corporaciones, instituciones u organizaciones, como es el caso del Área de Investigación en Sistemas de Información en donde de manera constante se dan a conocer notas, proyectos, conferencias relacionadas con el Área de Sistemas de Información y esto sirve a los usuarios para estar informados en el ámbito tecnológico o si estuvieran interesados en las actividades realizadas dentro de la Universidad Autónoma Metropolitana Unidad Azcapotzalco, como pueden ser conferencias, seminarios, etc.

Figura 5. Opción CONTACTOS del sistema web.

En la Figura 6, se puede visualizar el menú que está conformado por seis botones los cuales corresponden a los patrones lingüísticos, estos botones al ser activados muestran el resultado que contiene el patrón lingüístico que es elegido por el usuario. En el Capítulo 4, se mostrarán pruebas de dichos resultados al ser activado este menú.

Figura 6. Menú para elegir opciones de eventos indicadores en el sistema web.

3.3.2 Funcionamiento del sistema web

A continuación, se describen algunas actividades que se realizaron para llegar al funcionamiento correcto del sistema web del proyecto.

- Una vez que se obtuvieron los archivos que contiene los patrones lingüísticos a visualizar, la siguiente tarea fue leer dichos archivos para así guardarlos en un arreglo y posteriormente tener acceso a cada una de las líneas que conforman el archivo.

- Ya que se tiene el arreglo, se generó un número aleatorio que puede tomar el valor de 0 hasta el tamaño del arreglo generado.

- Teniendo el número aleatorio este sirvió para buscar dentro del arreglo y así tener acceso a la línea que indica el número obtenido.

- Consiguiendo la línea a leer esta se manda al jsp para poder visualizarla en el sistema web y así poder ver una línea distinta cada vez que se pulse el botón del cual se desea leer algún resultado.

Para la implementación y funcionamiento correcto del menú de resultados fue necesario agregar las líneas de código que se muestran en la Figura 7. En el Apéndice B se muestra el código completo y comentado.

```
        while(l!=null) //este ciclo while se usa para repetir el proces
        {
            l=br.readLine();//leemos una linea de texto y la guardamos
            String aux;/*variable auxiliar*/
            aux=l;/*si la variable l tiene datos se va acumulando en la

            arregloString.add(aux); /*se añaden al arregloString los el
        }
    }

      int numero;

      numero = (int) (Math.random() * (arregloString.size()));

      salida=( arregloString.get(numero));


}catch(IOException e){
    System.out.println("Error:"+e.getMessage());
}
return salida;
```

Figura 7. Código para el despliegue de resultados.

CAPÍTULO 4. Resultados

El objetivo de este capítulo es mostrar el funcionamiento correcto del sistema web, realizando algunas pruebas con los seis eventos indicadores que son representados por el menú de botones de lado izquierdo del sistema web.

4.1 Resultados

4.1.1 Pruebas que contienen el patrón lingüístico "… NUM por ciento …"

Para llevar a cabo las pruebas del sistema, se realizaron cinco pruebas con las diferentes opciones que se identificaron de los eventos indicadores. En las Figuras 8, 9, 10, 11 y 12 se muestran pruebas de resultados obtenidos que contienen el evento indicador % y que se mostrarán mediante el botón "porcentaje", que se encuentra en la parte izquierda del menú de opciones de los eventos indicadores.

Figura 8. Prueba uno del evento indicador % en el sistema web.

Figura 9. Prueba dos del evento indicador % en el sistema web.

Figura 10. Prueba tres del evento indicador % en el sistema web.

Figura 11. Prueba cuatro del evento indicador % en el sistema web.

Figura 12. Prueba cinco del evento indicador % en el sistema web.

4.1.2 Pruebas que contienen el patrón lingüístico "… NUM por ciento …"

Para llevar a cabo las pruebas del sistema, se realizaron cinco pruebas con las diferentes opciones que se identificaron de los eventos indicadores. En las Figuras 13,14,15,16 y 17 se muestran pruebas de resultados obtenidos que contienen el evento indicador por ciento y que se mostrarán mediante el botón "por ciento", que se encuentra en la parte izquierda del menú de opciones de los eventos indicadores.

Figura 13. Prueba uno del evento indicador por ciento en el sistema web.

Figura 14. Prueba dos del evento indicador por ciento en el sistema web.

Figura 15. Prueba tres del evento indicador por ciento en el sistema web.

Figura 16. Prueba cuatro del evento indicador por ciento en el sistema web.

Figura 17. Prueba cinco del evento indicador por ciento en el sistema web.

4.1.3 Pruebas que contienen el patrón lingüístico "… la mitad …"

Para llevar a cabo las pruebas del sistema, se realizaron cinco pruebas con las diferentes opciones que se identificaron de los eventos indicadores. En las Figuras 18,19,20,21 y 22 se muestran pruebas de resultados obtenidos que contienen el evento indicador la mitad y que se mostrarán mediante el botón "la mitad", que se encuentra en la parte izquierda del menú de opciones de los eventos indicadores.

Figura 18. Prueba uno del evento indicador la mitad en el sistema web.

Figura 19. Prueba dos del evento indicador la mitad en el sistema web.

Figura 20. Prueba tres del evento indicador la mitad en el sistema web.

Figura 21. Prueba cuatro del evento indicador la mitad en el sistema web.

Figura 22. Prueba cinco del evento indicador la mitad en el sistema web.

4.1.4 Pruebas que contienen el patrón lingüístico "… un tercio …"

Para llevar a cabo las pruebas del sistema, se realizaron cinco pruebas con las diferentes opciones que se identificaron de los eventos indicadores. En las Figuras 23,24,25,26 y 27 se muestran pruebas de resultados obtenidos que contienen el evento indicador un tercio y que se mostrarán mediante el botón "un tercio", que se encuentra en la parte izquierda del menú de opciones de los eventos indicadores.

Figura 23. Prueba uno del evento indicador un tercio en el sistema web.

Figura 24. Prueba dos del evento indicador un tercio en el sistema web.

Figura 25. Prueba tres del evento indicador un tercio en el sistema web.

Figura 26. Prueba cuatro del evento indicador un tercio en el sistema web.

Figura 27. Prueba cinco del evento indicador un tercio en el sistema web.

4.1.5 Pruebas que contienen el patrón lingüístico "… una cuarta parte …"

Para llevar a cabo las pruebas del sistema, se realizaron cinco pruebas con las diferentes opciones que se identificaron de los eventos indicadores. En las Figuras 28,29,30,31 y 32 se muestran pruebas de resultados obtenidos que contienen el evento indicador una cuarta parte y que se mostrarán mediante el botón "una cuarta parte", que se encuentra en la parte izquierda del menú de opciones de los eventos indicadores.

Figura 28. Prueba uno del evento indicador una cuarta parte en el sistema web.

Figura 29. Prueba dos del evento indicador una cuarta parte en el sistema web.

Figura 30. Prueba tres del evento indicador una cuarta parte en el sistema web.

Figura 31. Prueba cuatro del evento indicador una cuarta parte en el sistema web.

Figura 32. Prueba cinco del evento indicador una cuarta parte en el sistema web.

4.1.6 Pruebas que contienen el patrón lingüístico "… un total de …"

Para llevar a cabo las pruebas del sistema, se realizaron cinco pruebas con las diferentes opciones que se identificaron de los eventos indicadores. En las Figuras 33,34,35,36 y 37 se muestran pruebas de resultados obtenidos que contienen el evento indicador un total de y que se mostrarán mediante el botón "un total de", que se encuentra en la parte izquierda del menú de opciones de los eventos indicadores.

Figura 33. Prueba uno del evento indicador un total de en el sistema web.

Figura 34. Prueba dos del evento indicador un total de en el sistema web.

Figura 35. Prueba tres del evento indicador un total de en el sistema web.

Figura 36. Prueba cuatro del evento indicador un total de en el sistema web.

Figura 37. Prueba cinco del evento indicador un total de en el sistema web.

4.2 Análisis y Resultados

En esta sección se analizarán los resultados más específicos dentro del sistema web. Para llevar a cabo este proyecto se utilizaron alrededor de dos millones de noticias de las cuales se extrajeron más de 24 millones de oraciones, las cuales permitieron obtener aproximadamente quinientos mil resultados que contienen por lo menos un evento indicador que puede ser: "%"," por ciento"," la mitad"," un tercio"," una cuarta parte", "un total de".

¿Por qué decir que por lo menos un evento indicador? esto es porque por la forma en que se extrajeron las oraciones en un resultado pueden aparecer dos o más eventos indicadores como se puede apreciar en el ejemplo (13).

> (13) *Respecto a la población total de jóvenes que se tiene en México, el 2000 Instituto reveló que según las cifras del 2012 se tienen* ***un total de*** *31 millones de jóvenes en todo el país, lo cual refiere el* ***26.5%*** *de la población total.*

Cabe mencionar que este resultado se visualiza en las dos opciones del menú, la primera que es "porcentaje" y la última que lleva por nombre "un total de", ya que cada evento indicador proporciona información diferente.

También es importante resaltar que debido a que se manipularon notas periodísticas y cada persona es diferente al escribir estas, se obtuvieron resultados variados en cuanto al número de palabras que se pueden visualizar, ya que en promedio se calculó que los resultados contienen treinta y tres palabras, esto es porque si se reducen más las frases pierden el sentido informativo.

Después de realizar las pruebas necesarias para probar cada botón del menú en el sistema web, se puede observar que cada una de ellas fueron correctas ya que en todas se cumplió la función de mostrar resultados relacionados con el evento indicador elegido.

Conclusiones

Para el desarrollo de este Proyecto se tuvieron algunos problemas, en un principio debido a que el corpus es bastante grande, ya que contiene alrededor de dos millones de noticias, y había algunos limitantes con las herramientas a utilizar. Una de esas limitantes fue la memoria de la computadora que se utilizó para procesar el corpus.

Debido a lo anterior se perdió un poco de tiempo para poder generar los archivos que se visualizarían en el sistema web, sin embargo, después de varias pruebas se obtuvieron los archivos necesarios para mostrar los resultados.

Otro problema fue al visualizar los resultados ya que no se podían leer bien los acentos y la letra ñ, se probaron distintas soluciones que no resolvían el problema debido a que se creía que no estaban los archivos con la codificación UTF-8, que es un es un formato de codificación de caracteres, pero al final se pudo solucionar el problema leyendo los archivos de otra forma.

A pesar de todas las dificultades que se presentaron se cumplió con el objetivo principal, que era la visualización de un sistema web para la extracción y visualización de eventos indicadores a partir de un corpus de notas periodísticas en español utilizando patrones lingüísticos.

Una cosa que se puede implementar a futuro es la opción de compartir los resultados en las redes sociales como podría ser Facebook o Twitter, ya que como se mencionó anteriormente las redes sociales actualmente son un medio importante de comunicación.

Por último, debido a la experiencia y ayuda de mis asesores se tuvo la oportunidad de desarrollar y publicar un artículo el cual fue presentado en el *5th International Symposium on Language & Knowledge Engineering*, que se llevó a cabo en la Facultad de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla (BUAP), debido a que fue un foro para compartir nuevos conocimientos de investigación en el tema de procesamiento del lenguaje y otras áreas relacionadas, se pudo realizar esta publicación, ya que en este proyecto se abordó una de las tareas del Procesamiento del Lenguaje Natural que es la extracción automática de eventos.

Referencias bibliográficas

[1] Real Academia Española. (2001). Diccionario de la lengua española [*Dictionary of the Spanish Language*] (22nd ed.). Madrid, Spain: Author.

[2] F. J. Martín, J. L. Ruiz, "Procesamiento del lenguaje natural" [En línea]. España: Universidad de Sevilla, 2012-2013 Disponible en: https://www.cs.us.es/cursos/ia2/temas/tema-06.pdf

[3] L. D. Hernández, "Sistema web para identificar eventos y actores en textos periodísticos", proyecto terminal, Dep. de Sistemas, Universidad Autónoma Metropolitana Unidad Azcapotzalco, México, 2015.

[4] G. Moncecchi, A. Rosá, "Reconocimiento automático de eventos en textos", Proyecto de grado, Facultad de Ingeniería, Universidad de la República, Uruguay, 2010.

[5] L. Gil-Vallejo, I. Castellón, M. Coll-Florit, "Hacia una definición de la similitud verbal para la extracción de eventos", Centro Virtual Cervantes, 2015.

[6] A. Pinilla, M. Gutiérrez, L Ballejos, "Extracción automática de metadatos a partir de objetos de aprendizaje en un repositorio institucional: estado del arte". XLIII Jornadas Argentinas de Informática e Investigación Operativa (43JAIIO)-I Simposio Argentino de Tecnología y Sociedad (STS), ISSN: 2362-5139, pp. 67-82, 2014.

[7] M. Andersen, J. Hayes, A. Huettner, L. Schmandt, I. Nirenburg, "*Automatic Extraction of Facts from Press Releases to Generate News Stories", Proceedings of the Third Conference on Applied Natural Language Processing, Association for Computational Linguistics*, ANLC '92, pp. 170-177, 1992.

[8] TES (*Terminology Extraction Suite*): Distibución para Windows|Traducció, Traduccio.blogs.uoc.edu, http://traduccio.blogs.uoc.edu/2012/04/13/52/

[9] B. Priego-Sánchez, D. Pinto. "*Identification of Verbal Phraseological Units in Mexican News Stories*". Computación y Sistemas, Vol 19(4), pp. 713-720, 2015.

[10] O. Ramos, D. Pinto, B. Priego-Sánchez, I. Olmos, B. Beltrán, "Análisis empírico de la dispersión del español mexicano". *Research in Computing Science* 74, pp. 9-19, 2014.

[11] L. Padró, E. Stanilovsky, FreeLing 3.0: *Towards Wider Multilinguality Proceedings of the Language Resources and Evaluation Conference* (LREC 2012) ELRA.Istanbul, Turkey. May, 2012.

[12] Bienvenido a NetBeans y www.netbeans.org, Portal del IDE Java de Código Abierto, Netbeans.org, https://netbeans.org/index_es.html.

[13] ¿Qué es Java y para qué es necesario?,Java.com, https://www.java.com/es/download/faq/whatis_java.xml

[14] JavaScript, *Mozilla Developer Network*, https://developer.mozilla.org/es/docs/Web/JavaScript

[15] HTML, CSS3 y JQuery: curso práctico (2017). Ciudad de México: RA-MA Editorial, pp.52,93.

[16] 5 *Reasons to Learn HTML and CSS | SoloLearn: Learn to code for FREE!,*Sololearn.com, https://www.sololearn.com/Blog/12/5-reasons-to-learn-html-and-css/

Apéndices

Apéndice A. Publicación de Articulo en Congreso

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

LANGUAGE & KNOWLEDGE ENGINEERING LAB

BUAP

Awards this

CERTIFICATE to

Ariatna Quinto, Belem Priego-Sanchez, David Pinto and Jose Alejandro Reyes Ortiz

who presented the paper entitled

" Extracción automática de eventos indicadores a partir de noticias en español "

In the 5th International Symposium on Language & Knowledge Engineering

Puebla, Mexico, November 22nd - 24th, 2017

David Pinto, Ph.D.
Language & Knowledge Engineering Lab,
BUAP

Extracción automática de eventos indicadores a partir de noticias en español

Ariatna Quinto[1], Belém Priego Sánchez[1], David Pinto[2], José A. Reyes-Ortiz[1]

[1]Departamento de Sistemas

Universidad Autónoma Metropolitana unidad Azcapotzalco

[2]Facultad de Ciencias de la Computación Benemérita
Universidad Autónoma de Puebla

{al2113034676,abps,jaro}@azc.uam.mx,dpinto@cs.buap.mx

Resumen. Actualmente, se ha notado el incremento exorbitante de la informacion electrónica, claro ejemplo es la Web, la cual se ha convertido de fácil acceso. Ésta es posible estudiarla para saber los fenómenos que suceden a nivel de la lengua y a partir de ella se puede extraer meta informacion. En este artículo se han seleccionado artículos periodísticos en formato electrónico, como corpus, para llevar a cabo la extracción automatica de eventos indicadores que representan la proporcionalidad de un evento con respecto a un total, es decir, porcentajes relacionados a algún suceso. El principal objetivo de esta investigación es mostrar estadísticas que ocurren alrededor del mundo mediante la búsqueda de patrones lingüísticos en un sistema de recuperación de información. Este artículo presenta los avances obtenidos hasta el momento, en la extrac- cion automatica de eventos indicadores a partir de noticias en español.

Palabras clave: Extracción automática, recuperación de información, eventos indicadores.

Automatic Extraction of Indicative Events from Spanish News Stories

Abstract. Nowadays there exist an increasing raising of information on Internet which is easy available for human beings. This huge volume of data can be analyzed in order to discover and model linguistic pheno- mena and extract information. In this paper we have selected the news stories genre for extracting indicative events that represent likelihood of some event to occur. The aim of this research is to bring to light statis- tical events occurring all around the world by employing techniques of natural language processing based on linguistic patterns in an informa- tion retrieval system. This paper presents the outcomes obtained up to now in the particular topic of automatic extraction of indicative events from Spanish news stories.

Keywords. Automatic extraction, information retrieval, indicative events.

1. Introducción

Procesamiento de Lenguaje Natural (denotado por PLN), es una disciplina de la Inteligencia Artificial que trata la formulación e investigación de mecanismos de computación para la comunicación entre personas y máquinas, mediante el uso de Lenguajes Naturales. Dichos lenguajes son utilizados para la comunicación ya sea de forma escrita, hablada o en forma de signos [4]. Entre las tareas que realiza el PLN se encuentra la extracción automática de eventos, cuyo objetivo es capturar ciertas partes relevantes de un texto.

En el análisis del lenguaje se estudia la estructura del lenguaje a cuatro niveles [4]:

- Análisis morfológico: El análisis de las palabras para extraer raíces, rasgos flexivos, unidades léxicas compuestas y otros fenómenos.
- Análisis sintáctico. El análisis de la estructura sintáctica de la frase mediante una gramática de la lengua en cuestión.
- Análisis semántico. La extracción del significado (o posibles significados) de la frase.
- Análisis pragmático. El análisis de los significados más allá de los límites de la frase, por ejemplo, para determinar los antecedentes referenciales de los pronombres.

Este artículo, pretende abordar la tarea de la extracción automática de even- tos. En particular, los eventos serán los indicadores que representan la propor- cionalidad de un evento con respecto a un total que aparecen en una colección de información (noticias) en español, es decir, éstos son los porcentajes relacio- nados con algún suceso. Para el desarrollo total de esta investigación se hará uso únicamente de tres niveles del lenguaje, el morfológico (etiquetado POS), el sintáctico (segmentación del párrafo en oraciones) y el semántico (patrones lingüísticos). Además, se ha seleccionado el género periodístico debido a que es un tipo de escritura estándar y homogénea, en otras palabras, cualquier hablante nativo que lea el contenido de una nota periodística, lo entiende. Inclusive, la mayoría de las personas tiene acceso a un periódico, en formato papel o digital. Sin embargo, dado que es una investigación que está iniciando en este artículo, se presentarán los avances obtenidos al tratar de extraer de manera automática eventos indicadores a partir de noticias en español.

2. Motivación

Debido a la basta cantidad de información que actualmente se encuentra en la Web, ésta se puede utilizar y procesar de modo que se pueda emplear para ciertas tareas del PLN, una de ellas es la extracción de información relevante de un texto. Además de que es posible extraer conocimiento de toda esta información. Uno de los medios de comunicación que proporciona información es el Periódico, que especialmente en los últimos años su acceso ha sido en formato digital; esto debido a los avances tecnológicos, como es internet. Razón por la cual, en esta investigación se ha decidido trabajar con este género textual.

A través de los periódicos se ha podido llegar incluso a más gente y mantener un ritmo de actualización de los datos mucho más intenso que antes, siendo hoy imposible esperar de un día para otro para conocer noticias. Lo interesante de los periódicos es que cuando se habla de una sociedad, más o menos compleja, se pueden encontrar distintos tipos de periódicos que dan con el perfil de grupos sociales particulares, de grupos de edad, de regiones geográficas, de actividades laborales, de intereses específicos como deportes, internacionales, espectáculos o política.

Debido a que la sociedad cada vez vive de una manera más acelerada, se ha decidido extraer información de las noticias y a través de esta extracción dar a conocer datos estadísticos que la búsqueda de eventos indicadores proporcionará como resultado. Lo que

servirá para informar de manera más concisa y directa cierto suceso, esta información representará la proporcionalidad de un evento, sin la necesidad de que las personas lean toda la nota periodística, o más notas, para conocer datos importantes y relevantes. Al finalizar la investigación, se pretende crear un sistema web capaz de mostrar los resultados de la extracción de eventos indicadores. Dicho sistema servirá como medio de información resumida, lo cual beneficiará a la sociedad que desea estar informada.

3. Descripción del problema

En el Diccionario de la Real Academia [9], una de las acepciones para "evento" dice que es un suceso importante. De esta manera, este artículo retoma esa idea para mostrar un suceso importante que surge alrededor del mundo a través de los relatos periodísticos que se encuentran en la web. Sin embargo, se sabe que dar una definición concisa de evento es difícil. Por lo tanto, en esta investigación se tratará a un suceso como cualquier tipo de situación o acontecimiento que ocurre, restringiendo a eventos relacionados con la proporcionalidad de uno con respecto a un total, es decir porcentajes relacionados a un suceso. En (1) y (2) se pueden observar enunciados que muestran eventos indicadores.

(1) En México, cerca del 88 % de la energía primaria que se consume proviene del petróleo.

(2) La mitad de la población mundial está concentrada en tan solo seis países.

En (1) el evento indicador de porcentaje está explícitamente determinado por su signo ortográfico. Mientras que en (2) se nota un evento indicador, la mitad de la población mundial, pero no está explícitamente denotado por un signo ortográfico. Por lo que la tarea de extracción de eventos indicadores se convierte más complicada a medida que el lenguaje se desarrolla con diferentes factores de pensamiento, es decir, a medida que el lenguaje es abundante, enriquecido y creativo, la tarea se complica.

Típicamente, los eventos pueden ser expresados por verbos conjugados o en infinitivo, predicados en general y frases preposicionales. Sin embargo, como se ha notado, en esta investigación no se cumple con esos acuerdos, ya que se pueden encontrar eventos expresados mediante un signo de puntuación (%), partes que dividen un todo (mitad, tercio, etc.), entre otras expresiones. Sobre un texto plano, se pretende identificar eventos indicadores mediante diferentes patrones lingüísticos que serán identificados y son los que ayudaran a la extracción de diferentes sucesos acontecidos alrededor del mundo. Se puede delimitar entre las etiquetas <*indicador*>y <*/indicador*>, el inicio y fin del indicador porcentual con el fin de poder extraer toda la idea que contiene al evento, en (3) se puede observar un ejemplo. Esto mediante un sistema que segmente a los párrafos en oraciones.

(3) La <*indicador*>mitad<*/indicador*> de la población mundial está con- centrada en tan solo seis países.

En el contexto de este trabajo se contará con un esquema que establece reglas claras con guías de cómo se deben identificar los eventos indicadores a partir de diferentes patrones lingüísticos, con el objetivo de reducir las ambigëdades al mínimo.

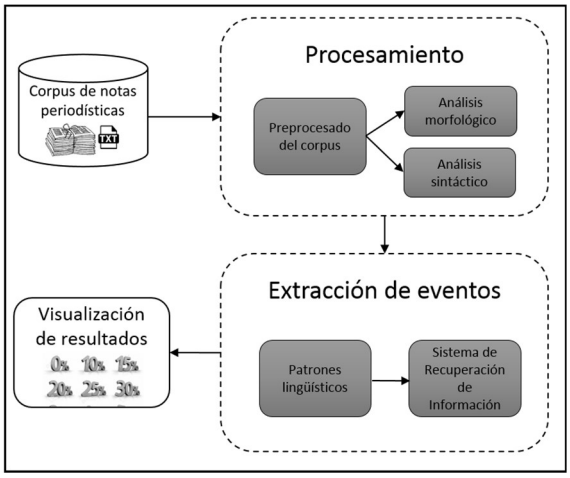
4. Estado del arte

En esta sección, se describen los trabajos reportados en la literatura relaciona- dos con la extracción automática de eventos. Cabe mencionar, que se incluyen trabajos que extraen eventos, sin embargo, corresponden a otro tipo que difiere al presentado en este trabajo, pero se incluyen debido a la temática presentada. La utilización de característicasicas sintácticas, semánticas y contextuales, mi- nería de textos, es uno de los varios enfoques que se utilizan para la extracción de eventos. El trabajo de Hernández [3], aborda su investigación sobre dicho enfoque, diseñó e implementó un sistema web para la anotación semántica de actores y eventos a partir de un corpus de textos periodísticos mexicanos. Este trabajo tiene la

particularidad principal, que la extracción la realiza en textos periodísticos mexicanos, sin embargo, son relatos periodísticos completamente diferentes a los que se abordan en este trabajo. El trabajo de Gil-Vallejo et al. [2], incluido en este enfoque, extrae eventos mediante las entidades y rela- ciones que posee el texto usando información lingística, relaciones sintácticas y semánticas, obteniendo un patrón de extracción de información relevante para un cierto evento. Los eventos que considera son los que contienen verbos dentro de su estructura de frase.Las técnicas de aprendizaje automático, también, son consideradas para el reconocimiento de eventos. Un trabajo presentado por Moncecchi y Rosá [5], que entra en el marco de este enfoque, reconoce eventos en textos españoles utilizando como base el esquema de anotación SIBILA [14] y dos algoritmos de aprendizaje automático, campo aleatorio condicional (CRF, por sus siglas en inglés Condi- tional Random Fields) [11] y máquinas de soporte vectorial (SVM, por sus siglas en inglés Support Vector Machines) [10], logrando mejores resultados con SVM. Además de investigaciones, existe el desarrollo de herramientas capaces de extraer eventos en un texto. JASPER [1] y TES [12], son un claro ejemplo de ello. En la primera, JASPER: Journalist's Assistant for Preparing Earnings Re- ports [1], se extraen ciertas piezas clave de información de un rango limitado de texto. El sistema está basado en el uso de plantillas, técnicas de compren- sión parcial y procedimientos heurísticos para extracción de información. Esta información, puede ser utilizada de varias maneras, como rellenar valores en una base de datos, generar resúmenes del texto de entrada, entre otras. Los eventos que principalmente extrae JASPER son los comunicados de prensa para generar historias de noticias. Con respecto a la segunda, TES: Terminology Extraction Suite [13], es una herramienta desarrollada para la extracción automática de terminología, que permite obtener términos y buscar automáticamente equiva- lentes de traducción. La herramienta está escrita en Perl, con interfaces gráficas implementadas en Tk. En este caso, los términos son eventos o sucesos que un documento, texto, contiene.

5. Metodología propuesta

En la extracción automática de eventos indicadores a partir de noticias en es- pañol, se han identificado dos etapas principales (ver Figura 1), las cuales serán alimentadas por un corpus de notas periodísticas en texto plano escritas en es- pañol,. A partir de estas notas se desarrollará la extracción automática de los eventos indicadores. La extracción se llevará a cabo mediante el diseño de patro- nes lingüísticos que cubrirán tres niveles de la lengua, morfológico, sintáctico y semántico. Finalmente, se pretende visualizar los resultados en un sistema web.



. **Figura 1.** Metodología propuesta para la extracción automática de eventos indicadores

5.1. Conjunto de datos

En esta sección se describe el conjunto de datos, corpus periodístico, en español de notas periodísticas que será utilizado para la extracción automática de eventos indicadores. Cabe mencionar que la descripción realizada es general debido a que el corpus a utilizar es el realizado en [9].El corpus ha sido extraído del sitio de intenet de la Organización Editorial Mexicana[1] , OEM, que contiene relatos periodísticos escritos en español me- xicano. A pesar de ser un sitio mexicano, no excluye las notas periodísticas internacionales pero de igual manera escritas en español mexicano. Los relatos periodísticos corresponden al período de tiempo del año 2007 al 2013.

Si bien, el corpus presenta diferentes metadatos, para el caso de este artículo, sólo se considerará el texto plano de la nota periodística. El corpus utilizado para esta tarea consta de 378,890 noticias, un total de 4,579,284 oraciones y alrededor de 11,1595,71 palabras.

5.2. Etapa de procesamiento

La etapa de procesamiento comprende el preprocesado del corpus, el análisis morfológico y sintáctico de dicho conjunto de datos.

La primera actividad, el preprocesado del corpus, contempla la eliminación de signos de puntuación y de palabras cerradas, mediante la utilización de un lexicón de éstos [8]. Además, se eliminan los caracteres especiales, poniendo atención a los caracteres relacionados con los eventos indicadores (%); estos caracteres a eliminar, se consideran irrelevantes para la extracción de los e entos indicadores debido a que ocasionarán conflicto al momento de recuperar la información que se está extrayendo y podrían incrementar el tiempo de respuesta.

La segunda actividad, el análisis morfológico, realiza el etiquetado de las partes de la oración (PoS de sus siglas en inglés, Part of Speech) en las notas periodísticas; se hace uso de las herramientas FreeLing [6] y/o TreeTagger[13]. El etiquetado consiste en identificar la categoría gramatical de cada palabra y asignarle una etiqueta dependiendo de la categoría gramatical a la que corres- ponda.

La tercera actividad, el análisis sintáctico, busca segmentar los párrafos que componen al corpus periodístico en oraciones, debido a que será más accesible la manipulación de oraciones y éstas tendrán una longitud más regular, es decir, el tamaño de un párrafo tiene menos proporcionalidad con respecto a una oración.

5.3. Etapa de extracción de eventos

La etapa de extracción de eventos incluye dos actividades esenciales, el diseño de los patrones lingísticos, para la extracción automática de eventos indicadores, y la búsqueda de éstos en un sistema de recuperación de información (denotado por SRI).

[1] Para mas informacion sobre la Organización Editorial Mexicana consultar: https://www.oem.com.mx/oem/

La primera actividad, diseño de patrones lingísticos, permite realizar el diseño de los patrones ling´ısticos que, en una etapa posterior, se implementan con el fin de extraer los eventos indicadores. Este diseño de patrones permite descubrir los elementos lingüísticos empleados con frecuencia en las notas periodísticas, para ello se utiliza uno o varios modelos que sirven como muestra para identificar y agrupar los eventos indicadores.

La segunda actividad, búsqueda de los patrones lingísticos en un SRI, posibi- lita, una vez que se han identificado los patrones lingüísticos, la implementación de éstos. Es decir, mediante un SRI, alimentado con el corpus de notas periodísti- cas y como consulta los patrones lingüísticos identificados, extraer la información relevante relacionada a los eventos indicadores.

6. Resultados obtenidos

El proceso de extracción de patrones lingüísticos se ha llevado a cabo mediante una técnica conocida como bootstrapping. Se considera un conjunto inicial de muestras etiquetadas manualmente, las cuales son posteriormente enriquecidas usando muestras similares obtenidas mediante un sistema de recuperación de información. De esta manera, es posible obtener un conjunto considerable de da- tos que comparten una estructura morfosintáctica que permite obtener patrones lingísticos que muestran la regularidad de estructuras para un tipo de expresión ling´ıstica en particular.

En la Tabla 1 se muestran ejemplos de los patrones lingüísticos más comunes encontrados mediante este proceso de generalización basado en un conjunto ini- cial moderadamente pequeño de muestras manualmente etiquetadas, pero que fue enriquecido mediante la técnica anteriormente mencionada.

Tabla 1. Patrones lingüísticos más frecuentes para eventos indicadores usando explíci- tamente el porcentaje.

Patrón lingüístico	**Semántica**
NUM por ciento	Indicador puntual alrededor del
NUM por ciento	Indicador aproximado mayor
que NUM por ciento	Indicador de punto base entre
NUM y NUM por ciento	Intervalo
de NUM a NUM por ciento	Incremento
hasta un NUM por ciento	Máximo temporal
NUM por ciento mas que	Indicador comparativo incremental un
incremento de NUM por ciento	Indicador comparativo incremental NUM por ciento
menos que	Indicador comparativo decremental un
decremento de NUM por ciento	Indicador comparativo decremental

Es importante aclarar que la etiqueta **NUM** se refiere a la especificación de un número en el texto periodístico en cualquiera de sus expresiones. Ejemplos de **NUM** ser´ıan los siguientes: 90, 28.3, noventa y tres, etc.

Por otro lado, en la Tabla 1, se ha usado el texto “por ciento”, el cual puede ser encontrado también sustituido por el símbolo ortográfico “ %” por lo que el lector debe considerar que los patrones anteriormente mencionados pueden ocurrir con cualquiera de estas dos expresiones textuales.

En la Tabla 2, se muestran otros patrones morfosintácticos que expresan el uso de números fraccionarios porcentuales, pero expresados en lenguaje natural.

Tabla 2. Patrones lingüísticos más frecuentes para eventos indicadores que usan otro tipo de expresiones del lenguaje natural relacionadas implícitamente con un porcentaje.

Patrón lingüístico	Semántica
un total de NUM de las/los NUM	Indicador porcentual basado en cociente la mitad Indicador 50 %
un tercio	Indicador 33 %
una cuarta parte	Indicador 25 %

Ejemplos de oraciones que contienen a algunos de los patrones presentados se muestran en la Tabla 3.

Tabla 3. Ejemplos de eventos indicadores.

Oraciones del género periodístico con un evento indicador
Una cuarta parte de hogares poblanos apenas tiene acceso a Internet
Sólo un tercio -31.8 por ciento- de los poblanos consideró al gobierno municipal Juntos canjearon 39 planillas, la mitad con nombre de Eugenio, la otra ...
Ventas de comercio al por menor crecerán hasta 7 por ciento en 2016
La economía informal contribuyó con el 24.8 por ciento del Producto Interno Bruto Esto supone entre un 25 y un 30 por ciento de los ingresos
Nuevo Sistema de Justicia Penal, al 90 % en Michoacán
El 75 % de los trabajadores en México está sometido a algún grado de estrés laboral, y eso a la larga es la causa del 25 % de los 75 mil infartos ...

7. Conclusiones y perspectivas

En este trabajo se han presentado experimentos relacionados con la extracción de eventos indicadores que utilizan un número fraccionario (tomando como base el 100) para expresar una unidad de valor.

El trabajo aporta una serie de patrones morfosintácticos útiles en la tarea de identificación de eventos indicadores. Se han extraído y presentado aquellos patrones lingüísticos que han mostrado una mayor regularidad de ocurrencia en los eventos indicadores.

Como trabajo a futuro se considera incrementar sustancialmente el número de noticias sobre el cual se llevarán a cabo los experimentos y llevando a cabo una

evaluación manual de todos y cada uno de los eventos extraídos manualmente, lo cual será por supuesto una tarea costosa desde el punto de vista del tiempo y esfuerzo humano.

Es importante analizar el conjunto inicial de eventos indicadores, a fin de poder enriquecer los patrones morfosintácticos y encontrar otros que aunque poco frecuentes, sean de interés en la extracción de la información basada en eventos.

Referencias

1. Andersen M., Hayes J., Huettner A., Schmandt L., Nirenburg I.: Automatic Ex- traction of Facts from Press Releases to Generate News Stories, Proceedings of the Third Conference on Applied Natural Language Processing, Association for Compu- tational Linguistics, ANLC ̀92, pp. 170-177, (1992)
2. Gil-Vallejo L., Castellon I., Coll-Florit M.: Hacia una definición de la similitud verbal para la extracción de eventos, Centro Virtual Cervantes, (2015)
3. Hernández L. D.: Sistema web para identificar eventos y actores en textos periodísti- cos, royecto terminal, Departamento de Sistemas, Universidad Autónoma Metropo- litana Unidad Azcapotzalco, México, (2015)
4. Martín F. J., Ruiz J. L.: Procesamiento del lenguaje natural, España: Universidad de Sevilla. Disponible en: https://www.cs.us.es/cursos/ia2/temas/tema-06.pdf (2013)
5. Moncecchi G., Rosá A.: Reconocimiento automático de eventos en textos, Proyecto de grado, Facultad de Ingeniería, Universidad de la República, Uruguay, (2010)
6. Padr L., Stanilovsky L.: FreeLing 3.0: Towards Wider Multilinguality Procee- dings of the Language Resources and Evaluation Conference (LREC 2012) EL- RA.Istanbul, Turkey, (2012)
7. Priego-Sánchez B., Pinto D.: Identification of Verbal Phraseological Units in Mexi- can News Stories. Computación y Sistemas, Vol 19(4), pp. 713-720, (2015)
8. Ramos O., Pinto D., Priego-Sánchez B., Olmos I., Beltrán B.: Análisis empírico de la dispersión del español mexicano. Research in Computing Science 74, pp. 9-19, (2014)
9. Real Academia Española.: Diccionario de la lengua española [Dictionary of the Spa- nish Language] (22nd ed.). Madrid, Spain, (2001)
10. Steinwart I., Christmann A.: Support Vector Machines (1st ed.). Springer Publis- hing Company, Incorporated, (2208)
11. Sutton C., McCallum A.: An introduction to conditional random fields for relatio- nal learning, in: L. Getoor, B. Taskar (Eds.), Introduction to Statistical Relational Learning, Ch.1, MIT Press, (2007)
12. TES (Terminology Extraction Suite): Distibución para Windows—Traducció, Tra- duccio.blogs.uoc.edu, urlhttp://traduccio.blogs.uoc.edu/2012/04/13/52/
13. TreeTagger, urlhttp://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
14. Wonsever D., Malcuori M., Rosá-Furman A.: SIBILA: Esquema de anotación de eventos". Reportes Técnicos 08-11. UR. FI INCO, (2008)

Apéndice B. Palabras cerradas en Español

A	B	C	D	E	F	G	H
a	bajo	cabe	da	e	fin	gran	ha
acá	bastante	cada	dado	ejemplo	fue	grandes	haber
actualmente	bien	casi	dan	el	fuera	gueno	había
adelante	buen	cerca	dar	él	fueron		habían
afirmó	buena	cierta	de	ella	fui		habrá
agregó	buenas	ciertas	debe	ellas	fuimos		hace
ahí	bueno	cierto	deben	ello			haceis
ahora	buenos	ciertos	debido	ellos			hacen
ajena		cinco	decir	embargo			hacer
ajenas		comentó	dejar	empleais			hacerlo
ajeno		como	dejó	emplean			haces
ajenos		cómo	del	emplear			hacia
al		con	demas	empleas			haciendo
algo		conmigo	demás	empleo			hago
algún		conocer	demasiada	en			han
alguna		conseguimos	demasiadas	encima			hasta
algunas		conseguir	demasiado	encuentra			hay
alguno		considera	demasiados	entonces			haya
algunos		consideró	dentro	entre			he
allá		consigo	desde	era			hecho
alli		consigue	después	eramos			hemos
allí		consiguen	dice	eran			hicieron
alrededor		consigues	dicen	eras			hizo
ambos		contigo	dicho	eres			hoy
ampleamos		contra	dieron	es			hubo
ante		cosas	diferente	esa			
anterior		creo	diferentes	esas			
antes		cual	dijeron	ese			
añadió		cuales	dijo	eso			

apenas aproximadamente aquel aquella aquellas aquello aquellos aqui aquí arriba aseguró asi así atras aun aún aunque ayer		cualquier cualquiera cualquieras cuan cuán cuando cuanta cuánta cuantas cuántas cuanto cuánto cuantos cuántos cuatro cuenta	dio donde dos durante	esos esta está ésta estaba estaban estado estais estamos estan están estar estará estas éstas este éste esto estos éstos estoy estuvo etc ex existe existen explicó expresó			

I	J	L	M	N	O	P	Q
igual	jamás	la	manera	nada	o	para	que
incluso	junto	lado	manifestó	nadie	ocho	parece	qué
indicó	juntos	largo	mas	ni	os	parecer	quedó
informó		las	más	ningun	otra	parte	queremos
intenta		le	mayor	ningún	otras	partir	querer
intentáis		les	me	ninguna	otro	pasada	quien
intentamos		llegó	mediante	ningunas	otros	pasado	quién
intentan		lleva	mejor	ninguno		pero	quienes
intentar		llevar	mencionó	ningunos		pesar	quienesquiera
intentas		lo	menos	no		poca	quienquiera
intento		los	mi	nos		pocas	quiere
ir		luego	mia	nosotras		poco	quiza
		lugar	mía	nosotros		pocos	quizas
			mias	nuestra		podeis	
			mientras	nuestras		podemos	
			mio	nuestro		poder	
			mío	nuestros		podrá	
			mios	nueva		podrán	
			mis	nuevas		podria	
			misma	nuevo		podría	
			mismas	nuevos		podriais	
			mismo	nunca		podríamos	
			mismos			podrian	
			modo			podrían	
			momento			podrias	
			mucha			poner	
			muchas			por	
			muchísima			porque	
			muchísimas			posible	
			muchísimo			primer	
			muchísimos			primera	

			mucho muchos muy			primero primeros principalmente propia propias propio propios próximo próximos pudo pueda puede pueden puedo pues	

R	S	T	U	V	Y
realizado	sabe	tal	última	va	y
realizar	sabeis	tales	últimas	vais	ya
realizó	sabemos	tambien	ultimo	valor	yo
respecto	saben	también	último	vamos	
	saber	tampoco	últimos	van	
	sabes	tan	un	varias	
	se	tanta	una	varios	
	sea	tantas	unas	vaya	
	sean	tanto	uno	veces	
	segun	tantos	unos	ver	
	según	te	usa	verdad	
	segunda	tendrá	usais	verdadera	
	segundo	tendrán	usamos	verdadero	
	seis	teneis	usan	vez	
	señaló	tenemos	usar	vosotras	

	ser	tener	usas	vosotros	
	será	tenga	uso	voy	
	serán	tengo	usted	vuestra	
	sería	tenía	ustedes	vuestras	
	si	tenido		vuestro	
	sí	tercera		vuestros	
	sido	ti			
	siempre	tiempo			
	siendo	tiene			
	siete	tienen			
	sigue	toda			
	siguiente	todas			
	sin	todavía			
	sín	todo			
	sino	todos			
	so	tomar			
	sobre	total			
	sois	trabaja			
	sola	trabajais			
	solamente	trabajamos			
	solas	trabajan			
	solo	trabajar			
	sólo	trabajas			
	solos	trabajo			
	somos	tras			
	son	trata			
	soy	través			
	sr	tres			
	sra	tu			
	sres	tú			
	sta	tus			
	su	tuvo			

	sus suya suyas suyo suyos	tuya tuyo tuyos			

Apéndice C. Clase para leer archivo y obtener el resultado a visualizar

```
public class Archivo {
public String leer(String nombre){
    ArrayList <String> arregloString=new ArrayList();
    String salida=null;
//Lectura del archivo de texto
//El parametro nombre indica el nombre del archivo
    try{
        FileInputStream fstream= new FileInputStream(nombre);
        InputStreamReader Fichero=new InputStreamReader(fstream,"UTF-8");
        BufferedReader br=new BufferedReader(Fichero);
       //Esta variable "l" la utilizamos para guardar más adelante toda la lectura
       del archivo
        String l="";

//Este ciclo while se usa para repetir el proceso de lectura, ya que se lee solo 1
línea de texto a la vez
       while(l!=null)
            {
          l=br.readLine();//leemos una línea de texto y la guardamos en la variable l
          String aux; //variable auxiliar
          aux=l; //si la variable l tiene datos se va acumulando en la variable aux,
       en caso de ser nula quiere decir que ya hemos leído todo el archivo de texto
          arregloString.add(aux); //se añaden al arregloString los elementos que se
van leyendo
            }

int numero; //variable entera
//generar un numero aleatorio de 0 hasta el tamaño     del arregloString
numero = (int) (Math.random() * (arregloString.size()));
//lectura de la línea que corresponde al número aleatorio antes generado
salida=(arregloString.get(numero));

    } catch(IOException e){
        System.out.println("Error:"+e.getMessage()); /*excepción en caso de no leer
el archivo*/
    }
    return salida;  //retorno de la variable a leer en html
    }
}
```

Apéndice D. Código jsp para visualización de sistema web

```
<%@page import="lectura.*"%>
<%@page pageEncoding="UTF-8"%>
<!DOCTYPE html>
<html>
    <head>
        <title>Extracción Automática de Eventos Indicadores</title>
        <meta http-equiv="Content-Type" content="text/html" charset="Unicode" />
        <meta charset="UTF-8">
```

```
        <meta name="viewport" content="width=device-width, initial-scale=1.0">
        <link rel="stylesheet"  href="css/estilos.css">
        <link href="https://fonts.googleapis.com/css?family=Cinzel|Courgette|Satisfy"
   rel="stylesheet">
<link href="https://fonts.googleapis.com/css?family=Patrick+Hand"   rel="stylesheet">
<link href="https://fonts.googleapis.com/css?family=Amatic+SC|Vidaloka”
   rel="sylesheet">
<link href="https://fonts.googleapis.com/css?family=Special+Elite" rel="sylesheet">
        <link href="https://fonts.googleapis.com/css?family=Pacifico"
   rel="stylesheet">
    </head>
<script>
   function mostrarPestana(n) {
   var pestanas = document.getElementsByClassName("pestana");
   var cabecera = document.getElementsByClassName("cabecera");
   for (i = 0; i < pestanas.length; i++) {
     if (pestanas[i].className.includes("p-activa")) {
       pestanas[i].className pestanas[i].className.replace("p-activa", "");
       cabecera[i].className = cabecera[i].className.replace("c-activa","");
     break;
     }
   }
   pestanas[n].className += " p-activa";
   cabecera[n].className += " c-activa";
   }
</script>

    <body>
        <div id="contenedor">
         <header>
         <h1> Extracción Automática de Eventos Indicadores %</h1>
         </header>

   <div class="contenido">
   <ul>
   <li><a href="javascript:location.reload()" class="cabecera c-activa"
   onclick="mostrarPestana(0);">Inicio</a></li>
   <div class="pestana p-activa"></div>

   <li><a href="#" class="cabecera" onclick="mostrarPestana(1);">ACERCA DE</a> </li>
   <div class="pestana">
         <h1>ACERCA DE</h1>
         <p>Este sistema web permite visualizar e integrar los resultados obtenidos en
         el proceso de la extracción de eventos indicadores</p>
   </div>

   <li><a href="#" class="cabecera" onclick="mostrarPestana(2);">ELABORADO POR</a></li>
   <div class="pestana">
         <h1>Elaborado por:</h1>
         <p>Alumna: Ariatna Quinto Martínez</p>
         <p>Asesora: Dra. Angeles Belém Priego Sánchez</p>
         <p>Asesor: Dr. José Alejandro Reyes Ortíz</p>
   </div>

   <li><a href="#" class="cabecera" onclick="mostrarPestana(3);">CONTACTOS</a></li>
   <div class="pestana">
         <h1>Contactos</h1>
         <p>Ariatna Quinto Martínez:  ariatna_09@<a
```

```
        href=https://outlook.live.com/owa/?authRedirect=true
        target="_blank">hotmail.com</a></p>
        <p>Dra. Angeles Belém Priego Sánchez:  abps@<a
         href=https://nechikali.azc.uam.mx/webmail/login.php
        target="_blank">correo.azc.uam.mx</a></p>
        <p>Dr. José Alejandro Reyes Ortíz:  jaro@<a
        href="https://nechikali.azc.uam.mx/webmail/login.php"
        target="_blank">correo.azc.uam.mx</a></p>
        <p><a href="https://es-la.facebook.com/AISII.UAM/" target="_blank"><img
        src="img/fb.jpg" title="Facebook"></a>     <a
        href="https://twitter.com/aisii_uam?lang=es" target="_black"><img
        src="img/twitter.png" title="Twitter"></a></p>
   </div>

   <img src="img/medios.jpg">
</div>
</ul>

   <section>
         <h1>¿Sabías que?</h1>
         <br>
  <div class='scroll'>
         <div class="porcentaje" id="porcentaje" style='display:none;'>
               <% Archivo a = new Archivo();
               String salida1 = a.leer("src/txt/resultado_1.txt");%>
               <p><%out.println(salida1); %><p>
         </div>

         <div class="porCiento" id="porCiento" style='display:none;'>
               <% Archivo b = new Archivo();
               String salida2 = b.leer("src/txt/resultado_2.txt");%>
               <p><%out.println(salida2); %></p>
         </div>

         <div class="laMitad" id="laMitad" style='display:none;'>
               <% Archivo c = new Archivo();
               String salida3 = c.leer("src/txt/resultado_3.txt");%>
               <p><%out.println(salida3); %></p>
         </div>

         <div class="unTercio" id="unTercio" style='display:none;'>
               <% Archivo d = new Archivo();
               String salida4 = d.leer("src/txt/resultado_4.txt");%>
               <p><%out.println(salida4); %></p>
               </div>

         <div class="unaCuartaParte" id="unaCuartaParte"     style='display:none;'>
               <% Archivo e = new Archivo();
               String salida5 = e.leer("src/txt/resultado_5.txt");%>
               <p><%out.println(salida5); %></p>
         </div>

         <div class="unTotalde" id="unTotalDe" style='display:none;'>
               <% Archivo f = new Archivo();
               String salida6 = f.leer("src/txt/resultado_6.txt");%>
               <p><%out.println(salida6);%></p>
         </div>
   </div>
```

```
                    <br>
                    <br>
                </section>
                   <br>
                   <br>
                <aside>
                    <br>
                <input type="submit" value="porcentaje"  name="porcentaje"
                onclick="cambiaVisibilidadPorcentaje()">
                    <br>
                    <br>
                <input type="submit" value="por ciento" name="por ciento"
                onclick="cambiaVisibilidadPorCiento()"/>
                    <br>
                    <br>
               <input type="submit" value="la mitad" name="la mitad"
                onclick="cambiaVisibilidadLaMitad()"/>
                    <br>
                    <br>
               <input type="submit" value="un tercio" name="un tercio"
                onclick="cambiaVisibilidadUntercio()"/>
                    <br>
                    <br>
               <input type="submit" value="una cuarta parte" name="una cuarta parte"
               onclick="cambiaVisibilidadUnaCuartaParte()"/>
                    <br>
                    <br>
               <input type="submit" value="un total de" name="un total de"
                onclick="cambiaVisibilidadUnTotalDe()"/>
                </aside>

<script type="text/javascript">
   function cambiaVisibilidadPorcentaje() {
   document.getElementById('porcentaje').style.display = 'block';
   document.getElementById('porCiento').style.display = 'none';
   document.getElementById('laMitad').style.display = 'none';
   document.getElementById('unTercio').style.display = 'none';
  document.getElementById('unaCuartaParte').style.display = 'none';
  document.getElementById('unTotalDe').style.display = 'none';
  }

   function cambiaVisibilidadPorCiento() {
   document.getElementById('porCiento').style.display = 'block';
   document.getElementById('porcentaje').style.display = 'none';
   document.getElementById('laMitad').style.display = 'none';
   document.getElementById('unTercio').style.display = 'none';
   document.getElementById('unaCuartaParte').style.display = 'none';
   document.getElementById('unTotalDe').style.display = 'none';
   }

   function cambiaVisibilidadLaMitad() {
   document.getElementById('laMitad').style.display = 'block';
   document.getElementById('porCiento').style.display = 'none';
   document.getElementById('porcentaje').style.display = 'none';
   document.getElementById('unTercio').style.display = 'none';
   document.getElementById('unaCuartaParte').style.display = 'none';
   document.getElementById('unTotalDe').style.display = 'none';
  }
```

```
  function cambiaVisibilidadUntercio() {
  document.getElementById('unTercio').style.display = 'block';
  document.getElementById('laMitad').style.display = 'none';
  document.getElementById('porCiento').style.display = 'none';
  document.getElementById('porcentaje').style.display = 'none';
  document.getElementById('unaCuartaParte').style.display = 'none';
  document.getElementById('unTotalDe').style.display = 'none';
  }

  function cambiaVisibilidadUnaCuartaParte() {
  document.getElementById('unaCuartaParte').style.display = 'block';
  document.getElementById('porCiento').style.display = 'none';
  document.getElementById('porcentaje').style.display = 'none';
  document.getElementById('laMitad').style.display = 'none';
  document.getElementById('unTercio').style.display = 'none';
  document.getElementById('unTotalDe').style.display = 'none';
  }

  function cambiaVisibilidadUnTotalDe() {
  document.getElementById('unTotalDe').style.display = 'block';
  document.getElementById('unaCuartaParte').style.display = 'none';
  document.getElementById('porCiento').style.display = 'none';
  document.getElementById('porcentaje').style.display = 'none';
  document.getElementById('laMitad').style.display = 'none';
  document.getElementById('unTercio').style.display = 'none';
  }

  </script>
  <footer>&copy;Extracción Automática de Eventos Indicadores -Todos los derechos
   Reservados-</footer>
            </div>
    </body>
</html>
```