Tratamiento automático de textos en español

Luis Villaseñor, Aurelio López,
Manuel Montes
Laboratorio de Tecnologías del Lenguaje, INAOE
Claudia Vázquez
Facultad de Filosofía y Letras, BUAP

The overwhelming presence of computers in our daily life has changed our way of working and communicating among us. This, along with the emergence of Internet -that enormous net that ties together all the world- has given birth to the so-called society of information. This society is based on the capacity of communication and distribution of information that offers that worldwide net. Of course, the central element of this communication is the human language and basically the available resources are, in their great majority, texts, that is, documents in a written form. The present work describes the efforts that our group carries out in the field of the automatic processing of written documents. For this aim, we introduce the situation of the Spanish language in the world of Internet; we place our work in relation to other language technologies; and finally we describe the processes involved in this automatic processing.

Palabras claves: *tecnologías del lenguaje, tratamiento del lenguaje natural, extracción de información, búsqueda de información, español de México.*
Fecha de recepción del manuscrito: marzo del 2003

Luis Villaseñor, Aurelio López y Manuel Montes.
Laboratorio de Tecnologías del Lenguaje del Instituto Nacional de Astrofísica, Óptica y Electrónica.
Luis Enrique Erro No. 1, Sta. María Tonantzintla, Puebla, México
correos electrónicos: villasen@inaoep.mx, allopez@inaoep.mx *y* mmontesg@inaoep.mx.

Claudia Vázquez
Facultad de Filosofía y Letras de la Benemérita Universidad Autónoma de Puebla
4 sur, No. 104, Centro,Puebla, México
correo electrónico: civonne22@hotmail.com.

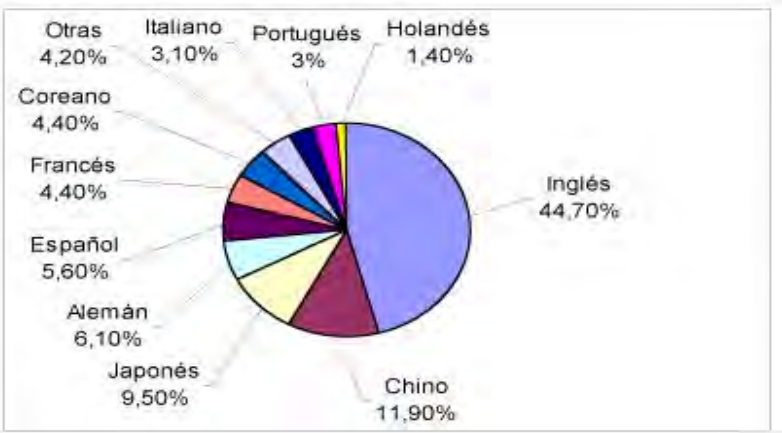
La avasallante presencia de las computadoras en nuestra vida diaria ha cambiado nuestra forma de trabajar y de comunicarnos. Aunada a la aparición de Internet –esa enorme red que enlaza a todo el mundo– ha nacido la llamada sociedad de la información. Esta sociedad está cimentada en la capacidad de comunicación y distribución de información que nos brinda esa red mundial. Por supuesto, el elemento central de esta comunicación es el lenguaje humano y básicamente los recursos disponibles son, en su gran mayoría, textos, es decir, documentos en forma escrita. El presente trabajo describe los esfuerzos que nuestro grupo realiza en el campo del tratamiento automático de documentos escritos. Para ello, presentamos la situación del idioma español en el mundo de Internet; ubicamos nuestro trabajo con relación a otras tecnologías del lenguaje; y finalmente describimos los procesos en que consiste este tratamiento automático.

1. Antecedentes

En nuestros días, debido a los desarrollos en medios de comunicación y de almacenamiento, existe más información disponible de la que somos capaces de leer, ya no digamos de analizar con suficiente detalle para darle un uso específico. Tras la aparición de Internet y de otros soportes electrónicos, millones de personas alrededor del mundo comparten diariamente grandes volúmenes de información. Lo anterior nos lleva a una situación en la cual dicho volumen de información crece día a día, lo que nos impide tener una idea global sobre la información relacionada con algún problema. Lo cotidiano es que debamos hacer juicios o tomar decisiones con la información parcial y fragmentada con la que contamos.

Aún cuando diversos tipos de información están disponibles en la actualidad, uno de ellos sigue predominando, el lenguaje escrito. Es decir, el lenguaje escrito continúa siendo un elemento clave en la llamada sociedad de la información.

Figura 1. Los lenguajes usados en Internet al 2001 (Martín 2000)



Las ciencias y las tecnologías que nos han permitido tener como nunca tanta información disponible, no han resuelto por completo los problemas asociados a la selección, búsqueda y análisis de la misma. En particular, las bases de datos, una de las áreas de investigación en computación en que más se ha trabajado y que tiene que ver con el manejo de "información" estructurada, no resuelve los problemas. Lo anterior debido a que, por un lado se encargan de manejar, como su nombre lo indica, datos, es decir no administran información. Por otro lado, como se ha observado, la información se encuentra principalmente expresada como lenguaje escrito (texto), con todas las complejidades que esto implica para su manejo y acceso.

1.1. El idioma español en la sociedad de la información

El idioma español es hablado por aproximadamente 400 millones de personas, incluyendo todas sus variedades fonéticas. Sin embargo, la influencia hoy en día de nuestro idioma en la sociedad de la información es mínima, como puede observarse en la figura 1. Esta situación se estima que irá gradualmente cambiando. La figura 2 nos muestra una proyección al 2005 de la penetración que tendrán las principales lenguas en Internet (Marcos 2000).

Figura 2. Evolución del número de usuarios de Internet (en millones)

Idioma	2000	2001 E	2005 E	Población Total	Penetración en 2005
Español	21	28	85	332	26%
Japonés	39	48	105	125	84%
Alemán	22	30	58	98	59%
Francés	17	22	38	72	53%
Chino	31	60	250	885	28%
Portugués	11	15	40	170	24%
Otros	25	81	132	-	-
Inglés	192.9	225	320	500	64%
Total no ingleses	211	278	820	5780	15%
Total	**391**	**503**	**1140**	**6085**	**18%**

Desgraciadamente la ciencia y, más concretamente, la tecnología asociada al tratamiento del idioma español no se ha desarrollado con el ímpetu necesario. Nos encontramos con un enorme rezago tecnológico que sólo podrá resolverse a través de la participación decidida del mundo hispanohablante. Defender una mayor presencia de contenidos propios en español no tiene por qué ser un reclamo de tipo nacionalista. En realidad se trata de una cuestión de supervivencia económica y cultural absolutamente legítima que afecta a todos los países hispanohablantes. Si queremos crecer en las redes y sacar partido de una comunidad de 400 millones de personas, no sólo es importante incrementar los contenidos en español, también es indispensable crear herramientas propias capaces de tratar la información expresada en nuestro idioma. Es por ello indispensable el desarrollo de una infraestructura tecnológica para el tratamiento y la interpretación automática de la información que se exprese en lenguaje español.

Cabe hacer mención de la responsabilidad de México en la búsqueda de soluciones a esta problemática. México es el país hispano hablante más poblado con cerca de 100 millones de habitantes sin contar la enorme presencia de los emigrantes mexicanos en E. U. A., una comunidad de más de 20 millones de personas.

El desarrollo de las tecnologías del lenguaje permitirá acceder, gestionar, intercambiar y analizar la información contenida en documentos digitales (textos, videos, grabaciones) escritos o hablados en español. Este desarrollo es un proyecto interdisciplinario que deberá integrar el trabajo, experiencia y conocimiento de los lingüistas mexicanos con el trabajo, experiencia y conocimiento de especialistas en ciencias de la computación.

2. Las tecnologías del lenguaje

Las tecnologías del lenguaje son el conjunto de conocimientos y medios involucrados en el tratamiento automático del medio de transmisión de información más complejo de nuestro planeta: el lenguaje humano (Cole *et al.* 1996). El lenguaje humano existe tanto en forma oral como escrita. Mientras que la forma oral es el modo de comunicación más antiguo y natural, la forma escrita es usada para conservar y transmitir el conocimiento humano. Las tecnologías del lenguaje, de habla y texto, procesan o producen expresiones en estos dos tipos de formas del lenguaje. A pesar de esta división, el lenguaje tiene aspectos que son compartidos entre el habla y el texto tal como los diccionarios, la gramática, significado de las oraciones, etc. Por otro lado una gran parte de las tecnologías del lenguaje no pueden reducirse únicamente a las tecnologías del habla o del texto. Entre esas tecnologías encontramos las que ligan el lenguaje al conocimiento. Nosotros no sabemos cómo el lenguaje, el conocimiento y el pensamiento están representados en el cerebro humano; sin embargo, las tecnologías del lenguaje proponen sistemas formales de representación que ligan el lenguaje a conceptos y tareas del mundo real. Además, el lenguaje humano incluye otros modos de comunicación. Por ejemplo, el habla se combina con ademanes indicativos y expresiones faciales; los textos digitales presentan combinaciones con imágenes y sonidos; una película además de la imagen puede contener lenguaje en forma oral o escrita, etc. De esta manera, las tecnologías del lenguaje incluyen muchas otras tecnologías que facilitan el procesamiento de la comunicación *multimodal* y de los documentos *multimedia*.

A continuación se describen brevemente algunas áreas de aplicación de dichos sistemas. Posteriormente se explican varias de las principales tecnologías del lenguaje que hacen posibles dichas aplicaciones.

2.1. Aplicaciones de las tecnologías del lenguaje

El propósito de las tecnologías del lenguaje es crear productos de software con algún grado de conocimiento del lenguaje humano, que permitan mejorar la interacción hombre-máquina. El desarrollo de dichos productos se basa en los siguientes planteamientos:

Las tecnologías amigables deben de escuchar y hablar. Uno de los objetivos de las tecnologías del lenguaje es la creación de modos de interacción más cercanos a la comunicación humana. Si un sistema automático es capaz de *conversar* con un ser humano

para solucionar una tarea, la herramienta se convertirá en asistente en la resolución del problema. Por supuesto, una conversación está restringida a un dominio específico pero nada impide tener numerosos asistentes, tantos como tareas existan. Ejemplos de este tipo de tareas son: la consulta de bases de datos (p. e. obtener información sobre la cartelera cinematográfica, el saldo de mi cuenta en el banco, etc.) o, el control de mecanismos (p. e. el control de una videocasetera o de una caldera).

Las máquinas deben facilitar la comunicación entre personas. Las tecnologías del lenguaje también ayudan a las personas a comunicarse entre sí independientemente de sus lenguas maternas. En este caso, el problema central, y que precisamente es uno de los objetivos iniciales de las tecnologías del lenguaje, es la traducción automática entre diferentes lenguajes. Actualmente sólo se tienen resultados modestos, pero que a pesar de ello ya son un gran soporte para los traductores humanos. La aplicación de las tecnologías de lenguaje relacionadas con la traducción automática tendrá un gran impacto social y económico. Por ejemplo, es bien sabido que el principal cuello de botella del comercio electrónico es la comunicación entre clientes y vendedores.

El lenguaje es la fábrica de la Web. El rápido crecimiento de Internet, acompañado del surgimiento de la sociedad de la información, estableció nuevos retos a las tecnologías del lenguaje. Básicamente, se requiere de software que permita navegar, filtrar y procesar el contenido de documentos *web*. Las tecnologías del lenguaje destinadas a estas tareas son y serán de gran importancia porque sólo a través de ellas, la información digital disponible en línea puede ser transformada en conocimiento colectivo. El contenido multilingüe de la web constituye un reto adicional para las tecnologías de lenguaje. Sólo los sistemas multilingües para la administración de información podrán traspasar las barreras del lenguaje para el comercio electrónico, educación y cooperación internacional.

2.2. Principales tecnologías del lenguaje

Los siguientes párrafos exponen brevemente el problema al que enfrentan las principales tecnologías del lenguaje.

Reconocimiento del habla. El lenguaje hablado (señal acústica) es reconocido y transformado a una representación escrita (texto).

Síntesis de voz. Se trata de la operación inversa del reconocimiento del habla. Las expresiones en lenguaje hablado son producidas a partir de texto (en sistemas texto a voz), o a partir de las representaciones internas de las oraciones.

Clasificación de textos. El objetivo es determinar automáticamente la clase o categoría temática de un texto. Esto se realiza a partir de un análisis léxico del texto, y del uso de conjuntos de textos de entrenamiento manualmente clasificados.

Generación automática de resúmenes. Aquí el objetivo es seleccionar las porciones más significativas de cada texto, y con ellas construir un resumen. La generación de resúmenes se complica cuando tiene que realizarse con base en una petición específica.

Búsqueda de información textual. En la también llamada recuperación de información, el problema a resolver es obtener los textos de una colección inicial que mejor casan con la petición dada por una persona con una necesidad de información y solo esos. Los documentos candidatos (recuperados) se ordenan basándose en su relevancia estimada.

Extracción de información. Las piezas de información predeterminadas en un texto son descubiertas y marcadas para su extracción. Estas piezas extraídas pueden ser: las fechas, nombres de lugares o personas, o relaciones tanto simples como complejas, como por ejemplo, precios de artículos o participantes en un accidente.

Sistemas de diálogo. En este caso el sistema puede sostener un diálogo con el usuario humano, en el cual, el usuario solicita información o realiza una compra, una reservación u otro tipo de transacción.

Traducción automática. Tecnologías que traducen textos o asisten a traductores humanos. Típicamente estas tecnologías usan grandes cantidades de textos, en conjunto con sus traducciones manuales de tal forma que sea factible hacer una adecuada traducción de palabras, frases y oraciones.

El carácter multidisciplinario en la búsqueda de soluciones en estas áreas es inherente. Para cada una de estas áreas es necesario contar con una gran cantidad de recursos y métodos elaborados en diferentes disciplinas. Éstos van desde métodos computacionales como: algoritmos diseñados específicamente para el análisis sintáctico; pasando por técnicas estadísticas, especialmente útiles en el reconocimiento del habla y la recuperación de información; hasta conocimientos lingüísticos indispensables para el *tratamiento* de los fenómenos del lenguaje.

3. Buscando información en textos en español

El trabajo realizado hasta ahora en nuestro laboratorio está principalmente relacionado al tratamiento de la información textual, en particular Búsqueda (BI) y Extracción de Información (EI). El objetivo es lograr que una persona con ayuda de una computadora pueda obtener reportes, no únicamente de textos previamente estructurados o de bases de datos expresamente diseñadas, sino de documentos escritos libremente en español. Para ello, es necesario dar a la computadora los elementos necesarios: (i) que le permitan identificar los documentos que discuten el tema en cuestión, y (ii) para identificar y obtener la información de interés de dichos documentos, con la creación resultante de una base de datos con la información extraída.

En el área de tratamiento de información en forma escrita ha habido desarrollos importantes para hacer filtrado, extracción, organización, búsqueda y análisis de la información. No obstante, estos avances han sido hechos principalmente para el inglés. Para el caso de información expresada en español existe un rezago importante. Las técnicas y herramientas existentes para otros idiomas no son inmediatamente aplicables a nuestro idioma. Este panorama nos enfrenta con la urgente necesidad de contribuir en la investigación y desarrollo de métodos y procedimientos para la gestión de documentos escritos en español.

Para ejemplificar lo anterior, presentamos un caso en particular: la búsqueda de información sobre desastres en nuestro país. Normalmente esta información es reportada a través de los diarios y hoy en día, gracias a Internet, podemos acceder a dichos reportes desde nuestra computadora personal. Sin embargo, hojear cada periódico para encontrar información específica demandaría de un esfuerzo enorme. Aún en el caso de buscar información sencilla y concreta, por ejemplo, cuándo y dónde tocó las costas mexicanas el huracán Isidore, requeriría el revisar varias notas periodísticas de varios días.

El laboratorio de Tecnologías del Lenguaje del INAOE trabaja actualmente en un proyecto que aborda esta problemática. La principal motivación para esto es que, como es sabido, nuestro país es especialmente vulnerable a incidentes que causan daños materiales y humanos. De ahí que propongamos alcanzar como un producto adicional de los esfuerzos de investigación, un almacén o repositorio de información de desastres que han ocurrido en el siglo XXI, desde su inicio hasta el período de vigencia del proyecto. Lo anterior junto con un sistema prototipo que ponga en práctica los logros del proyecto para estudios de prospección y prevención. La definición de los datos importantes a extraer de cada siniestro corresponde a los establecidos por la Red de Estudios Sociales en Prevención de Desastres en América Latina (www.desinventar.org). Por ejemplo, fecha y lugar de ocurrencia, duración y magnitud, y efectos tales como el número de muertos y damnificados, y el número de viviendas destruidas y dañadas, etc.

El primer paso involucra localizar y obtener notas como la siguiente de los periódicos accesibles a través de Internet:

> "El huracán Isidore perdió fuerza y fue clasificado como tormenta tropical, y dejó en la península de Yucatán 300 mil personas damnificadas y el deceso de una persona por imprudencia, pues pretendió hacer composturas eléctricas a la intemperie y se electrocutó, dio a conocer Carmen Segura, coordinadora general de Protección Civil de la Secretaría de Gobernación.
> "Indicó que el gobierno federal dispuso para los estados perjudicados una primera partida de 30 millones de pesos del Fondo Nacional de Desastres, y se enviaron a territorio yucateco tres contenedores con productos para atender la emergencia por vía terrestre, debido a las dificultades para hacerlo por aire.
> "A su paso por Yucatán, el meteoro dejó damnificadas a 65 mil personas, de las cuales 80 se refugiaron en 240 albergues, según reportes preliminares que ofreció el gobierno de esa entidad, mismo que atribuyó a Isidore tres fallecimientos.
> "Detalló que por el huracán se perdió en 60 por ciento la capacidad de ofrecer servicios de electricidad, agua potable y telefonía en Mérida, donde la tercera parte de las calles se inundaron."

A partir de esta nota se extraen datos específicos que describen el desastre (véase la tabla 1). Gracias a este tipo de fichas descriptivas se irá conformando una base de datos de desastres.

Tabla 1. Datos obtenidos de una nota periodística

Tipo de evento	**Huracán**
Personas afectadas	300 mil
Personas muertas	1
Lugar	Yucatán
Casas afectadas	
Casas destruidas	

Por otro lado, notas periodísticas como la anterior aportan información adicional describiendo circunstancias previas y posteriores al evento en cuestión. Dadas las circunstancias únicas de cada evento, la naturaleza de esta información complementaria es extremadamente diversa y, en consecuencia, difícilmente sintetizada en una tabla como la anterior. Por ejemplo, la información adicional podría describir "causas", "efectos" o la "respuesta que las autoridades o la sociedad" dieron a un evento. En la nota aludida tenemos que el evento "causó inundaciones en las calles" y el gobierno puso disponible "300 millones de pesos y tres contenedores de producto". Contrastando lo anterior con algún otro desastre natural, por ejemplo una explosión volcánica, los efectos y respuestas podrían ser completamente diferentes. Tratar automáticamente esta información *no estructurada* es un reto considerablemente complejo sobre el cual aún falta mucho por hacer. Para efectos de esta exposición, dejaremos de lado el tratamiento de información *no estructurada* y nos enfocaremos a la extracción automática de datos descriptivos determinados de antemano.

A continuación se discuten algunos problemas prevalecientes en la búsqueda de información y se detallan los conceptos fundamentales sobre los que descansa la extracción de información.

3.1. La búsqueda de información

Una persona con una necesidad de información puede aproximarse de distintas formas a una posible fuente, dependiendo de su nivel de conocimiento del dominio o del uso que le va a dar a la información. Así, puede ir de una actitud de curioseo, en la que usualmente no tiene el conocimiento o la claridad de su necesidad para formular una petición, a un planteamiento bien específico de su necesidad.

La búsqueda es la manera en que cualquier persona, por más casual que sea, confronta el caos que percibe al aproximarse a un sistema con millones de elementos de información y trata de entenderlos. Este es precisamente el caso que afrontamos cada vez que nos sentamos frente a nuestra computadora personal e intentamos buscar en la Web. Por supuesto, en la actualidad los sistemas automáticos de búsqueda facilitan enormemente esta tarea (Brewer 2001). Las respuestas de estos sistemas de búsqueda son sólo aproximaciones al documento que buscamos, pero el nivel de aproximación es asombroso si partimos del hecho que los sistemas de búsqueda "no entienden" los documentos, únicamente se basan en la presencia y frecuencia de los elementos léxicos por nosotros buscados.

En el contexto de nuestro interés, no es inmediato especificar "desastres" para obtener notas relevantes dado que muy pocas de las noticias detallando huracanes, erupciones, derrames químicos o inundaciones, hacen mención de que se trata de un desastre. Entonces, tendríamos que aproximar la máquina de búsqueda con una lista de palabras del tipo de desastres de interés como la listada anteriormente.

Aun suponiendo que únicamente estamos interesados en un sólo tipo de desastres, nos tenemos que enfrentar a una de las desventajas de las técnicas prevalecientes basadas en palabras clave o términos. Por ejemplo, al utilizar únicamente la palabra "huracán" en un buscador de Internet actual, seguramente obtendremos una enorme variedad de respuestas, éstas incluirán los documentos que discurren sobre la perturbación atmosférica, al igual que documentos que informan sobre un equipo de fútbol o publicitan una telenovela. Depender exclusivamente de palabras aisladas en la búsqueda de información siempre nos enfrentará al problema de la ambigüedad. Sin embargo, si se ofrece un contexto más amplio de los términos que se dan para búsqueda, por ejemplo "daños causados por huracán" o "club de fútbol Huracán", se eliminaría esta ambigüedad.

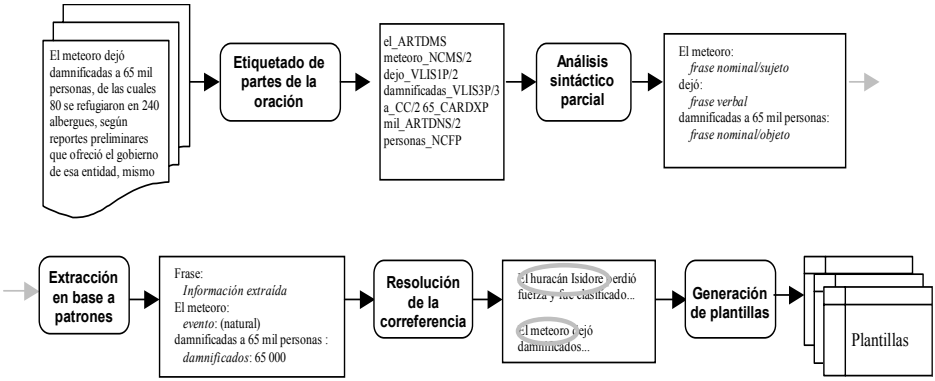
El uso de un contexto para hacer la búsqueda de información (Finkelstein *et al.* 2002) es una área que pretendemos investigar, principalmente en la información no estructurada. En esta dirección debe proponerse también una forma de integrar este uso con consultas tradicionales en bases de datos.

3.2. La extracción de información

El objetivo principal de la Extracción de Información (EI) es el procesamiento de textos escritos libremente con el fin de encontrar información útil con respecto a un dominio de interés predeterminado. La información *extraída* es entonces transformada a una representación fuertemente estructurada. En contraste con la búsqueda de información, la EI debe recorrer cada texto encontrando secciones relevantes para obtener la información útil (Grishman 1997). Esto implica que un sistema de este tipo debe conocer los mecanismos que estructuran el lenguaje escrito. Por supuesto, es imposible pensar en un sistema capaz de "comprender" cualquier tipo de texto. Por el momento, dos fuertes restricciones son aplicadas en este tipo de sistemas: (i) el dominio de interés está predeterminado de antemano, es decir, sólo extraeremos cierto tipo de información de textos previamente selec-

cionados; y (ii) el análisis sintáctico del texto se hace de manera parcial, enfocando nuestros esfuerzos en "patrones sintácticos" que sean los más probables a responder a nuestras necesidades de información.

Figura 3. Arquitectura general de un sistema de Extracción de Información



La figura 1 muestra la arquitectura clásica de un sistema de EI (Cardie 1997). Los párrafos subsecuentes describen brevemente los elementos centrales de esta arquitectura.

1) *Etiquetado en partes de la oración*. Durante esta fase se fragmenta el texto en unidades elementales —normalmente al nivel de palabras— y se asocia a cada unidad una etiqueta que describe su morfología y/o función gramatical en el texto.
2) *Análisis sintáctico parcial*. Esta etapa tiene por objetivo la identificación de frases nominales, frases verbales, frases preposicionales y otras estructuras sintácticas simples. Durante o posteriormente a este análisis se identifican entidades cuya función semántica es relevante al tema de interés. La diferencia principal entre este análisis parcial y un análisis sintáctico clásico radica en que no se desea construir un árbol sintáctico completo de cada frase del texto. El análisis sintáctico parcial sólo analiza fragmentos de texto que pueden ser reconocidos con un alto nivel de confianza.
3) *Extracción sobre la base de patrones*. Durante esta etapa el sistema identifica, a partir de un catálogo de patrones lingüísticos del tema en cuestión, los elementos relevantes. Estos elementos son extraídos y depositados en una plantilla descriptiva del evento.
4) *Resolución de la correferencia*. La correferencia es un fenómeno muy común en el texto escrito. Para determinar la mayor cantidad de elementos descriptivos de un evento dependemos de resolver las relaciones correctas entre las diferentes frases que refieren a éste. Fenómenos como la anáfora deben resolverse en este paso.

Aunque parezca sorprendente, la EI ya ha alcanzado niveles que la convierten en una tecnología viable para aplicaciones reales. Es cierto que aún falta mucho por recorrer para alcanzar niveles semejantes a analistas especializados en tareas de extracción de información. Pero, también es cierto, que la tarea de extracción de información para el ser humano común es una actividad difícil.

Por mencionar algunos ejemplos de la EI en el mundo real tenemos: i) el resumen de expedientes médicos extrayendo diagnósticos, síntomas, terapias, etc. (Soderland *et al.* 1995); ii) el análisis de transcripciones de cables informativos de radio y televisión para encontrar y resumir actividades terroristas (MUC-4 1992; MUC-3 1991); iii) la clasificación automática de documentos legales (Holowczak & Adam 1997). Por otro lado, un número creciente de aplicaciones para Internet usan la tecnología de EI. Algunos ejemplos incluyen: i) la creación de bases de datos sobre propuestas de empleo a partir de foros de discusión, portales electrónicos y anuncios clasificados; ii) la creación de bases de datos sobre información meteorológica a partir de páginas Web (Soderland 1997).

Como es de imaginar, los sistemas antes mencionados son para el tratamiento de textos en inglés. Los esfuerzos encaminados al tratamiento del español son pocos (Cardeñosa *et al.* 2000, Subirats & Ortega 2002).

4. Trabajo reciente en el laboratorio de tecnologías del lenguaje

A continuación presentamos detalles del trabajo realizado hasta ahora, en las áreas de búsqueda y extracción de información, dentro del contexto del proyecto "Recolección, Extracción, Búsqueda y Análisis de Información a partir de Textos en español" llevado a cabo en el laboratorio de Tecnologías del Lenguaje.

4.1. Recolección de información

El objetivo de esta primera etapa del proyecto es la construcción automática de un volumen de noticias sobre desastres naturales en México. Así pues, en ella nos enfocaremos básicamente en localizar y obtener notas que describan tanto el fenómeno natural como los daños ocasionados por éste. Un ejemplo de este tipo de notas es mostrada en párrafos anteriores (véase la sección 3).

Para llevar a cabo esta tarea son necesarios los siguientes dos puntos:

- Un sistema de navegación automática, capaz de monitorear sitios Web de periódicos en línea, para la recolección de notas periodísticas, y
- Un clasificador de texto que seleccione de entre toda la colección acumulada, las notas relevantes con información sobre algún tipo de desastre natural.

Para efectos de esta presentación enfocaremos nuestra atención sobre el segundo punto: el clasificador de textos.

4.1.1. La clasificación de textos

En los párrafos subsecuentes explicaremos uno de los experimentos realizados para determinar los criterios más adecuados para seleccionar notas relevantes al tema de desastres naturales. Cabe mencionar que por el momento, únicamente trabajaremos tres desastres naturales: huracán, inundación y sequía. Dado que esta operación de clasificación deberá ser realizada por una computadora, deseamos encontrar criterios de selección sencillos basados en la información léxica de las notas y tratar de evitar el arduo trabajo de analizar un texto para "comprender" su significado. Para lograr esto, fue necesario recopilar una colección para "entrenar" nuestro clasificador. El entrenamiento consiste en determinar automáticamente los elementos léxicos que mejor discriminan un texto relevante de uno irrelevante. Este conjunto de entrenamiento fue recolectado manualmente y se utilizó el periódico *Reforma* (www.reforma.com) como fuente de información inicial. De este sitio se recopilaron noticias relacionadas (tanto relevantes como irrelevantes) con los fenómenos naturales de huracán, inundación y sequía, correspondientes a los últimos dos años. Las noticias relevantes incluyen información del fenómeno natural, mientras que las catalogadas como irrelevantes contienen palabras o frases usadas comúnmente en la descripción de un fenómeno natural pero que se usan en contextos muy diferentes. Por ejemplo, la palabra huracán en el contexto de "el presidente está en el ojo del huracán".

El experimento descrito a continuación se basa en un conjunto de entrenamiento formado por 375 documentos, de los cuales el 11.5 % son noticias relevantes y el 88.5 % restante son irrelevantes. Básicamente, este experimento considera varias estrategias de extracción de características léxicas de los documentos, en particular, las estrategias de reducción de dimensionalidad conocidas como *umbral en la frecuencia* y *ganancia en la información*. Igualmente se consideraron dos métodos de clasificación: el método *simple de Bayes* y el de *vecinos más cercanos*.

4.1.2. Extracción de características

El primer paso en la búsqueda de criterios léxicos de clasificación es la caracterización de cada documento a partir de las palabras que encontramos en él. Por supuesto, no todas las palabras son elementos discriminantes, así el primer paso es la determinación de un conjunto de palabras o características léxicas pertinentes. Los pasos que se siguieron para determinar el conjunto de características más adecuado fueron: (i) pre-procesamiento, eliminando todas las marcas o vocablos irrelevantes, (ii) indexado de los documentos de nuestro corpus de entrenamiento, para determinar el número y frecuencia de los elementos léxicos, y (iii) reducción del conjunto de características a un número adecuado para mejorar los tiempos de cómputo, pero sin perder precisión en la capacidad de selección.

Pre-procesamiento. El propósito de esta etapa es reducir el tamaño de los documentos eliminando las partes de los textos que dan poca información sobre su contenido, es decir, que carecen de significado temático. El proceso realizado a cada uno de los documentos fue el siguiente:

- Eliminación de etiquetas HTML – debido a que las notas son recuperadas de portales Web, es necesario eliminar las etiquetas incrustadas en el documento que indican a un navegador como mostrarlo en pantalla. Por supuesto, estas etiquetas no proporcinan información útil en nuestra tarea de clasificación.
- Eliminación de símbolos de puntuación.
- Eliminación de palabras vacías. Estas son sobre todo aquellas partículas como prepo siciones o artículos.
- Reducción de palabras a su raíz. Por ejemplo "desconocer", "desconocerlos" y "desconocía" tienen la raíz léxica "desconoc".

La reducción en tamaño de cada documento fue en promedio aproximadamente del 52 % de su tamaño original.

Indexado. El proceso de indexado está basado en el modelo vectorial con pesado booleano. Durante este proceso se encontraron 310,498 instancias léxicas en el conjunto de entrenamiento, con un vocabulario de 29,710 palabras. Además la frecuencia de ocurrencia de los términos en el vocabulario varía entre 1 y 977.

Reducción de dimensionalidad. Desde el punto de vista computacional manejar un vocabulario de 29,710 palabras provocará problemas en los tiempos de respuesta. Basándonos en la idea intuitiva de que no todas estas palabras son necesarias para una correcta clasificación, aplicamos dos métodos para reducir este conjunto de características: basándonos en la frecuencia del término, y en la ganancia en la información (IG) (Sebastiani 1999). La elección de estos métodos se debe a que han revelado encontrarse entre los más efectivos (Yang y Pedersen 1997).

En los experimentos realizados se eliminaron los términos cuya frecuencia fue menor a diez ocurrencias, dando como resultado una reducción en el vocabulario, dejando sólo 2550 términos, y posteriormente se eliminaron los términos cuya IG fue cero. Es decir, sólo se tomaron los términos que dan información útil para la predicción de clases. El resultado fue un vocabulario de sólo 214 términos para el espacio de características, lo que refleja que solo el 0.7% del vocabulario es útil para la predicción de las clases. Los términos del vocabulario con mayor ganancia en la información fueron: meteorología (0.1327), tropical (0.1215), sequía (0.1105), viento (0.0974) y agua (0.0942).

4.1.3. Métodos de clasificación

Con base en los resultados reportados en la bibliografía reciente (Sebastiani 1999), se seleccionaron los métodos tradicionales de vecinos más cercanos y clasificador simple de Bayes. Ambos algoritmos han mostrado ser de los mejores en la tarea de clasificación de textos. El clasificador de vecinos más cercanos es un método basado en ejemplos o instancias, donde no se construye ninguna descripción de las categorías, más bien se utilizan directamente los ejemplos del conjunto de entrenamiento ya suminis-

trados para determinar la clasificación de ejemplos no vistos. Es decir, para decidir si un nuevo ejemplo pertenece a determinada categoría, se verifica si un cierto número de ejemplos ya dados y muy cercanos al nuevo ejemplo (vecinos), también pertenecen a la misma clase. El clasificador simple de Bayes es un método de tipo probabilístico que aplica, como su nombre lo indica, el teorema de Bayes con una suposición de independencia entre las coordenadas del vector representando los documentos. De esta forma, el objetivo de la clasificación es la estimación de los parámetros de una distribución de probabilidad que describa el conjunto de entrenamiento.

4.1.4. Resultados

Los resultados que se presentan en las siguientes tablas consideran los métodos de vecinos más cercanos y clasificador simple de Bayes. Asimismo analizan el efecto de aplicar reducción de dimensionalidad con las técnicas de umbral en la frecuencia y ganancia en la información al conjunto de características.

Para comparar y evaluar la efectividad de los clasificadores se usaron varios métodos. El primer método fue la validación cruzada con 10 subconjuntos *(10 Fold Cross Validation)* (Mitchell 1997; Witten y Frank 2000). En la tabla 2 se presentan los porcentajes de aciertos y fallos de cada clasificador usando validación cruzada.

Tabla 2. Evaluación a través de la validación cruzada

	Umbral en la frecuencia Frec > 10		**Umbral en la frecuencia Frec > 10 y Ganancia en la información IG > 0**	
	Vecinos más cercanos (K=1)	**Simple de Bayes**	**Vecinos más cercanos (K=1)**	**Simple de Bayes**
Instancias clasificadas correctamente	90.93 %	93.3 %	92.8 %	**97.06 %**
Instancias clasificadas incorrectamente	9.06 %	6.6 %	7.2 %	**2.93 %**

Como puede observarse en la tabla 2 los criterios para reducir la dimensionalidad resultaron adecuados; sin importar qué clasificador usemos los resultados fueron mejores al usar los dos criterios conjuntamente. También puede observarse en la tabla 2 que el clasificador basado en el método simple de Bayes es el que mejor resultados obtuvo en nuestro contexto, con una tasa de clasificación muy alentadora del 97 %.

La tabla 3 presenta la matriz de confusión de los resultados obtenidos para el mejor caso (clasificador simple de Bayes usando umbral en la frecuencia y ganancia en la información para reducción de la dimensionalidad). En esta matriz, la diagonal principal refleja las instancias clasificadas correctamente, y los valores fuera de la diagonal representan los documentos mal clasificados indicando a qué categoría erróneamente fueron asociados. Al observar estos datos, es interesante notar que el clasificador no confunde documentos relevantes entre ellos, es decir, no clasifica una nota sobre huracán en inundación. De esta manera podemos afirmar que no sólo la tasa de clasificación es muy alentadora, sino además, la calidad en la clasificación de los textos es notable pues los errores sólo se cometieron al considerar un documento relevante en la categoría *no relevante*.

Tabla 3. Matriz de confusión

	Huracán	**Inundación**	**Sequía**	**No relevante**
Huracán	15	0	0	3
Inundación	0	5	0	2
Sequía	0	0	14	4
No relevante	1	0	1	330

Otro método de evaluación de un clasificador de textos es usando las medidas de precisión y "recuerdo" (Lewis 1991). Ambas medidas son tradicionalmente usadas en la recuperación de información. La precisión expresa en que medida el clasificador toma una decisión correcta al ubicar cualquier documento en la clase que le corresponde. El "recuerdo" refleja cuántos de todos los documentos de una clase son clasificados en ella. Las tablas 4 y 5 muestran los resultados obtenidos con ambos clasificadores.

Podemos observar que el algoritmo simple de Bayes es mejor clasificando correctamente los documentos de cualquiera de las clases, cuando se toma en cuenta la ganancia de información. Asimismo, muestra un mejor "recuerdo" que los demás, es decir nos garantiza mejor la clasificación de los documentos de cualquiera de las clases.

Tabla 4. Evaluación de vecinos más cercanos

Umbral en la frecuencia Frec > 10		**Ganancia en la información IG > 0**		
Precisión	**"Recuerdo"**	**Precisión**	**"Recuerdo"**	**Clase**
0.889	0.444	0.846	0.611	Huracán
1	0.143	1	0.714	Inundación
0.667	0.111	0.6	0.333	Sequía
0.912	0.994	0.939	0.982	No relevante

Tabla 5. Evaluación de simple de Bayes

Umbral en la frecuencia Frec > 10		**Ganancia en la información IG > 0**		
Precisión	**"Recuerdo"**	**Precisión**	**"Recuerdo"**	**Clase**
1	0.278	0.938	0.833	Huracán
0	0	1	0.714	Inundación
0.933	0.778	0.933	0.778	Sequía
0.932	0.997	0.973	0.994	No relevante

4.2. Extracción de información

A diferencia de la clasificación de textos, en la EI es necesario hacer un análisis lingüístico más profundo de los documentos. Como se vio en la sección 3.1., es necesario hacer un análisis sintáctico *parcial*, así como un análisis para resolución de la correferencia. Nuestro enfoque difiere del tradicional al agregar una etapa inicial al esquema clásico de la EI. Básicamente esta nueva etapa consiste en la búsqueda de patrones léxicos.

Como se mencionó en la sección anterior, el análisis léxico es el más sencillo desde el punto de vista automático. Es por ello de gran interés la definición de mecanismos que exploten al máximo la información léxica dejando los menos puntos a resolver a través de los otros dos análisis subsecuentes. Los párrafos siguientes describen este nuevo mecanismo propuesto para el máximo aprovechamiento de la información léxica en la EI.

4.2.1. Un método de EI basado en técnicas de clasificación de textos

Se espera que éste método extraiga la mayor cantidad de información *interesante* de cada evento desastroso (fecha, lugar, duración, magnitud, número de muertos, etc.), usando únicamente información léxica. La idea de base de este método es la búsqueda automática de patrones léxicos que envuelven los datos que se desean extraer. Para encontrar estos patrones también serán usados clasificadores de texto. En este caso, en lugar de tener documentos relevantes y no relevantes, tendremos frases o segmentos de frases relevantes dado el dato que se desea extraer.

Para lograr esto debemos contar con un conjunto de entrenamiento, es decir, frases o segmentos de frases identificados como relevantes o irrelevante. El proceso de construcción de tal conjunto de entrenamiento consiste en identificar y anotar todos los datos deseados de un conjunto de documentos relevantes. Por supuesto, la anotación de estos textos debe realizarse de forma manual con criterios bien definidos. Un ejemplo de un texto anotado se muestra a continuación:

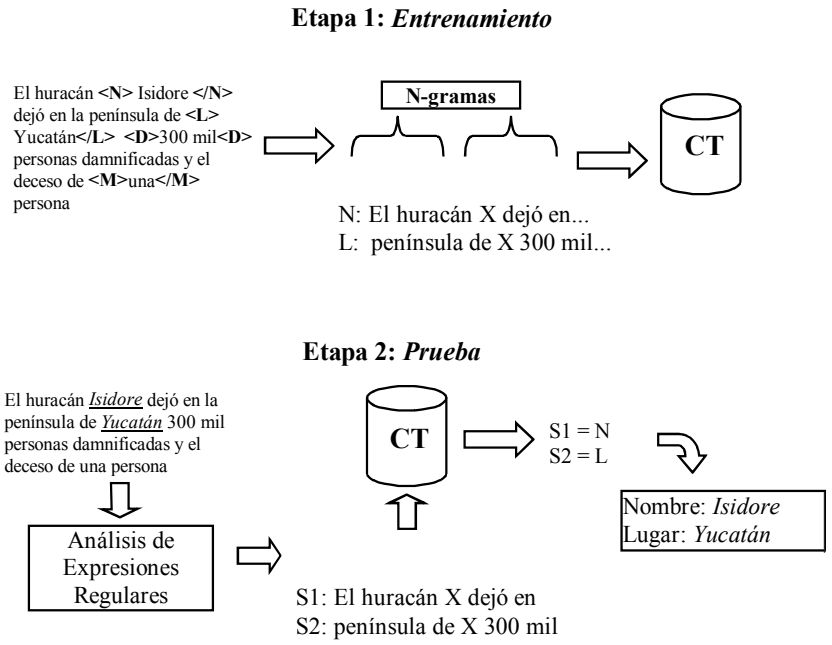
> El huracán **<N>** Isidore **</N>** dejó en la península de **<L>** Yucatán **</L>** **<D>** 300 mil **</D>** personas damnificadas y el deceso de **<M>** una **</M>** persona.

Cada una de la etiquetas encierra alguno de los datos deseados e indica la categoría a la que pertenece. Por ejemplo, en este caso tenemos la etiqueta <N> para indicar el nombre del fenómeno, <L> para el lugar, <D> para la cantidad de damnificados, y <M> para la cantidad de personas fallecidas.

A partir de este conjunto de entrenamiento etiquetado, el clasificador obtendrá todos los posibles patrones léxicos para cada una de las categorías. La figura 2 ilustra este método. Una vez obtenidos los criterios de selección estaremos en posibilidad de extraer la información deseada en nuevos documentos. Para ello primero se identifican, mediante la aplicación de expresiones regulares, los datos candidatos a ser extraídos (fechas, cantidades, nombres propios), y después el clasificador seleccionará los datos apropiados para finalmente almacenarlos en el campo correspondiente de la base de datos de salida.

Cabe resaltar que este método es de gran valía pues al ser automático encuentra de manera exhaustiva todos y cada uno de los patrones posibles. Por otro lado, el método tiene como gran inconveniente la necesidad de corpus anotados. Es claro que para enfrentar el problema de la extracción de información de manera integral son necesarios mecanismos para resolver fenómenos comunes en el lenguaje escrito, como por ejemplo, la anáfora. Por supuesto este método no pretende resolver este tipo de problemas. Sin embargo, para contextos sencillos con un lenguaje limitado o especializado, él podrá extraer el máximo de información de una manera relativamente sencilla.

Figura 4. Diagrama de **EI** basado en clasificadores de texto



5. El reto del tratamiento automático del lenguaje

Como puede observarse, la solución de la búsqueda y la extracción de información debe apoyarse en desarrollos tanto de la lingüística como de las ciencias computacionales. De hecho, la lingüística computacional es una disciplina que enlaza estos mundos y explora posibles soluciones al "entendimiento" del lenguaje por una computadora.

Tradicionalmente esta búsqueda ha explorado dos caminos. El primero utilizando teorías que intentan explicar cómo el ser humano comprende y usa el lenguaje; y el segundo concentrando sus esfuerzos —dada la complejidad del problema— en la identificación de patrones repetitivos presentes en el lenguaje sin aspirar a encontrar una explicación de su uso. Desde el punto de vista de la creación de sistemas automáticos, el primer camino es demasiado ambicioso siendo casi imposible desde el punto de vista computacional. El segundo camino reduce el lenguaje únicamente a la ocurrencia de secuencias de palabras. Esta propuesta puede implementarse fácilmente en un sistema computacional pero con resultados insuficientes. Actualmente, la línea divisoria entre ambos caminos es menos clara, dado que cada vez se proponen más mecanismos que mezclan y balancean estos dos enfoques.

Por ejemplo, en el caso de la extracción de información, durante el etiquetado en partes de la oración, un proceso estadístico decide sobre qué etiqueta asignar a cada palabra. Por supuesto, para ello fue necesario "entrenar" al *etiquetador*, es decir, alimentarlo con el mayor número de ejemplos posibles de cada una de las palabras del vocabulario deseado para calcular sus contextos de ocurrencia. Los ejemplos de muestra deben ser etiquetados por ojos expertos que determinan la etiqueta que mejor describe la función de la palabra en ese contexto. En el primer caso, tenemos un proceso que a partir de una tabla de probabilidades designa la etiqueta, no obstante, el cálculo de esas probabilidades se hizo a partir de recursos analizados desde un punto de vista lingüístico.

Esta historia se repite en otras ramas de las tecnologías del lenguaje. La construcción de un sistema automático empieza por un estudio lingüístico de un fenómeno en particular del lenguaje sobre un conjunto de textos escogidos. En cada ocasión que dicho fenómeno se presenta se etiqueta. Enseguida, aplicamos un proceso en la búsqueda de propiedades recurrentes que describan dicho fenómeno. A partir de estos patrones estamos en posibilidad de tratar el lenguaje de manera automática.

Hasta el momento, como se ha demostrado para otros lenguajes, es posible tratar de esta manera fenómenos sencillos del lenguaje, y es muy probable que los límites de esta técnica sean rápidamente alcanzados. Sin embargo, aun para este tratamiento "sencillo" queda mucho por hacer para el español.

6. Conclusiones

Hasta ahora esta breve exposición de la problemática que aborda el tratamiento del lenguaje escrito, es sólo un somero panorama de las enormes posibilidades y el gran reto a enfrentar en el tratamiento automático del lenguaje. Es clara la gran importancia y el enorme compromiso que tenemos para defender nuestro lenguaje y nuestra cultura en los medios electrónicos.

Es claro que las soluciones propuestas a este problema recaen en nuestro conocimiento de cómo el ser humano produce y comprende el lenguaje. Nuestro trabajo no pretende resolver esta gran pregunta, pero al proponer modelos que emulen mecanismos propios del lenguaje humano nos ayudará a descubrir y describir formalmente sus propiedades ocultas. A largo plazo, nuestra meta es la creación de poderosas aplicaciones con capacidades lingüísticas, ello favorecerá, por supuesto, el entendimiento del lenguaje humano.

Por otro lado, a pesar de que los resultados hasta ahora alcanzados están lejos de presentar la habilidad humana para el manejo del lenguaje, ellos son muy alentadores. Existen sistemas que demuestran su enorme utilidad, por ejemplo, los sistemas de dictado o los buscadores de información en la Web capaces de manipular enormes cantidades de documentos.

En resumen, esta tarea no es nada fácil y no podrá ser resuelta con algunos esfuerzos aislados. En este campo interdisciplinario es necesario trabajar conjuntamente, lingüistas e informáticos, compartiendo nuestras experiencias y conocimientos. Es por

eso de gran importancia unir nuestros esfuerzos tanto en la construcción de recursos lingüísticos, necesarios para ambas disciplinas, así como, de herramientas que sirvan, tanto para validar teorías lingüísticas que afronten problemas propios del español de México, como para la construcción de sistemas de uso práctico.

Referencias

BREWER E.A. (2001) "When everything is searchable", *Communications of the ACM*, March 2001, Vol. 44, No. 3: 53-55.

CARDEÑOSA, J., IRAOLA, L. & TOVAR, E. (2000) "Author extraction: a test experience for flexible information. Flexible query systems, recent advances". *Proceedings de la 4ª conferencia internacional sobre sistemas de consultas flexibles,* FQAS'2000, Physica-Verlag, 2000: 255—266.

CARDIE C. (1997) "Empirical methods in information extraction", *AI Magazine*, Winter 1997: 65-79.

COLE, R. A., MARIANI, J., USZKOREIT, H., ZAENEN, A. & ZUE, V. (1996) *Survey of the state of the art in human language technology*. 1996. http://cslu.cse.ogi.edu/HLTsurvey/

FINKELSTEIN L., GABRILOVICH E., MATIAS Y., RIVLIN E., SOLAN Z., WOLFMAN G. & RUPPIN E. (2002) "Placing search in context: the concept revisited", *ACM TOIS*, January 2002, Vol. 20 No. 1: 116-131.

GRISHMAN R. (1997) *Information extraction. Techniques and challenges*. Rome: Springer-Verlag, Lecture Notes in Artificial Intelligence.

HOLOWCZAK, R. & ADAM, N. (1997) "Information extraction-based multiple-category document classification for the global legal information network". In *Proceedings of the ninth conference on innovative applications of artificial intelligence*, Menlo Park, CA. AAAI: 1013-1018.

LEWIS, D. (1991) "Evaluating text categorization". *Proceedings of the speech and natural language workshop*, Asilomar, CA, Feb. 1991.

MARCOS MARÍN, F. A. (2000) "La lengua española en Internet". En *El español en el mundo.* Anuario 2000 del Instituto Cervantes. http://cvc.cervantes.es/obref/anuario/anuario_00/

MARTÍN MAYORGA, D. (2000) "El español en la sociedad de la información". En *El español en el mundo*. Anuario 2000 del Instituto Cervantes. http://cvc.cervantes.es/obref/anuario/anuario_00/

MITCHELL, TOM M. (1997) *Machine learning*. McGraw-Hill.

Muc-3 (1991) *Proceedings of the third message-understanding conference (MUC-3).* San Francisco, CA: Morgan Kaufmann.

Muc-4 (1992) *Proceedings of the fourth message-understanding conference (MUC-4).* San Francisco, CA: Morgan Kaufmann.

SEBASTIÁN, F. (1999) *Machine learning in automated text categorization: a survey*. Technical Report IEI-B4-31-1999. Istituto di Elaborazione dell'Informazione.

SODERLAND, S., ARONOW, D., FISHER, D., ASELTINE, J. & LEHNERT, W. (1995) *Machine learning of text-analysis rules for clinical records*. Technical report, TE39, Boston, Mass: Department of Computer Science, University of Massachusetts.

SODERLAND, S. (1997) "Learning to extract text-based information from World Wide Web". In *Proceedings of the third international conference on knowledge discovery and data mining*, 251-254. Menlo Park, CA: AAAI Press.

SUBIRATS, C, & ORTEGA, M. (2002). "EXTRACCIÓN AUTOMÁTICA DE INFORMACIÓN DE GRANDES CORPUS". En J. de Kock y C. Gómez (eds). *La lingüística de corpus: aplicaciones*. Salamanca: Ediciones Universidad de Salamanca.

WITTEN, Ian H. And FRANK, E. (2000) *data mining: practical machine learning tools and techniques with Java implementations*. Sydney: Morgan Kaufmann. 2000.

YANG, Y. And PEDERSEN, J. P. (1997) "Feature selection in statistical learning of text categorization". 14th International Conference on Machine Learning.