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

Propuesta de proyecto terminal:
Sistema de procesamiento de textos de investigación

Alumno:

Fernando Alejandro Acosta
208200575

Trimestre 2012 Otoño
10 de diciembre de 2012

Tercera versión

Asesor

Maricela Claudia Bravo Contreras
Profesor Asociado
Departamento de Sistemas

Objetivo general

Diseñar e implementar un sistema de anotado de textos basado en procesamiento de lenguaje natural para procesar y extraer información de textos de investigación digitalizados en inglés.

Objetivos específicos

- Diseñar e implementar un sistema de procesamiento de tratamiento sintáctico de textos.
- Diseñar e implementar un sistema para analizar semánticamente el contenido sintáctico de los textos.
- Diseñar una interfaz gráfica que permita seleccionar las etiquetas que serán agregadas al texto analizado.

Introducción

Desde 1940 el problema del almacenamiento y recuperación de la información ha tenido una atención especial, actualmente ha tomado gran importancia debido al acelerado desarrollo de los medios electrónicos para el almacenamiento de información. Un problema que ha surgido de forma paralela a este hecho lo encontramos cuando deseamos tener acceso a esta información y durante este proceso la precisión y la velocidad resulta no ser tan eficaz.

Realizar de forma manual la recuperación de la información tiene como consecuencia que en la mayoría de las veces parte de esta información es ignorada por la persona quien realiza este proceso principalmente por la falta de precisión necesaria que se requiere durante este trabajo.

Existe en la actualidad una gran cantidad de documentos de texto de investigación digitalizados como artículos y reportes técnicos que en su totalidad están escritos en el lenguaje humano, aquí es donde resulta necesario apoyarnos de las computadoras para explotar esta información, transformando esta información al lenguaje formal de la máquina.

Como respuesta a esta necesidad, hoy en día ha surgido de entre las ramas de la inteligencia artificial y la lingüística computacional, el procesamiento del lenguaje natural, que mediante la implementación de sus técnicas permite a las máquinas manejar lenguajes no formales.
En esta propuesta se desarrollará un sistema de procesamiento de textos de investigación para resolver el problema de realizar anotado de texto en documentos utilizando técnicas de tratamiento de textos del procesamiento del lenguaje natural.

Justificación

Se cuenta ya en la actualidad con una inmensa cantidad de documentos en distintos formatos que están digitalizados, obedeciendo a la vanguardia de la tecnología informática y apoyándonos de algunas herramientas que ya existen. Este proyecto pretende incorporarse a este campo que se encuentra en desarrollo y explotar el contenido de toda esta información de manera óptima.

Con el desarrollo de este proyecto se lograra la extracción automatizada de información relevante de un documento. Considerando el alcance del presente proyecto, se considera información relevante de los documentos de investigación a los siguientes datos: nombre de los autores, título del documento, filiación de los autores, correo electrónico palabras clave, resumen y nombre de la publicación. Dicha información será de gran utilidad en proyectos posteriores de minería de textos de investigación. Para el desarrollo de este proyecto se utilizaran técnicas recuperación de información y de procesamiento de lenguaje natural y de manera organizada los conocimientos de ingeniería.

Como resultado de llevar a cabo este proyecto se logrará un sistema capaz de identificar y etiquetar información relevante de documentos de investigación. En esta propuesta consideramos un documento de investigación a todo tipo de texto que reporta el resultado de alguna investigación o que es de divulgación científica. Los documentos de investigación son tradicionalmente publicados en: libros, tesis, reportes o artículos de congresos y de revistas. El proceso de anotado semántico agrega etiquetas que permiten identificar información relevante en los documentos que antes no estaban incluidas, como pueden ser etiquetas de <Autor>, <Correo>, <Título>, <Filiación>, <Palabras clave>, <Resumen>, etc. Las etiquetas no se agregaran sobre el mismo documento, estas etiquetas existirán como información almacenada y que esta asociada al mismo documento con la finalidad de que cuando esta información sea requerida esté disponible, con el propósito de facilitar tareas más complejas como: la búsqueda automatizada de documentos, la clasificación y agrupamiento de documentos, y la minería de textos y contenido.

Trabajos relacionados

En la universidad ya existen proyectos terminales con la que tiene relación:

- Sistema de recuperación de información semántico [1].

  El objetivo de este proyecto es la recuperación de información relevante sobre un conjunto de textos. La etapa inicial de este proyecto es la que mantiene una estrecha relación con este proyecto, ya que se aplican técnicas de procesamiento del lenguaje natural para realizar el análisis de los textos.
  También presenta una gran similitud en cuanto a los objetivos que incorpora a la universidad debido a que ambos proyectos pretenden crear una herramienta que recupera información relevante sobre las publicaciones de los profesores de la UAM Azcapotzalco.

  Aunque ambos proyectos tienen como punto de partida el procesamiento del lenguaje natural, difieren en el uso que se le da a la información recuperada mediante la construcción y uso de gráficas se representa el significado de la oración que después es convertida en información organizada y comprensible para el usuario.

- Clasificación de Servicios web semánticos mediante ontologías [2].

  Este proyecto mantiene una relación en cuanto a la necesidad que existe en ambos de diseñar e implementar un analizador sintáctico para extraer elementos relevantes, ambos sistemas parten de tener un parámetro de entrada que posteriormente es procesado.

  Una de las diferencias encontradas es que para este proyecto el analizador sintáctico es usado para extraer la información relevante pero de un servicio web semántico. Otra diferencia la encontramos en el uso que se le da a la información obtenida por el analizador sintáctico, en este proyecto es usado para clasificar.

- Sistema semántico para la búsqueda, selección y adecuación de contenidos educativos basados en perfiles de aprendizaje [3].

  Este proyecto tiene como objetivo el tratamiento de contenido educativo, encontramos relación en la forma en cómo se lleva a cabo este proceso, toma como punto de entrada una ontología de contenidos educativos y lo que realiza con esto es una búsqueda, que sería lo equivalente a un analizador sintáctico, posteriormente hace una clasificación que es parecido a un analizador semántico.

La diferencia la encontramos en la utilidad de la información que ha sido clasificada, en este proyecto es utilizado para recomendación y generación de estadísticas, mientras que en el sistema que se desarrollará será utilizada para realizar un etiquetado semántico.

Trabajos relacionados de fuentes externas de la universidad.

- Técnicas del Procesamiento del Lenguaje Natural [4].

  Este artículo resulta ser de especial interés debido a que proporciona una introducción de las características del procesamiento del lenguaje natural y la gran ayuda que brinda para la recuperación de la información, describe además técnicas de representación de documentos. Este artículo sirve como base para entender de forma más detallada el propósito que pretende solucionar este proyecto terminal.
  La diferencia encontrada radica principalmente en que este artículo explica de forma muy detallada en qué consiste el procesamiento del lenguaje natural y describe algunas técnicas existentes, pero nunca hace uso de ellas para resolver algún problema empleando estas herramientas, contrario a esta propuesta donde hacemos uso de estas técnicas para cumplir el objetivo de este proyecto.

- Sophia Semantic Engine [5].

  El motor de Sophia Semántica es un software comercial de origen Italiano que analiza y comprende el lenguaje natural, creando una capa de interpretación en las aplicaciones que interactúan con los usuarios de una forma lingüística y las aplicaciones que se ocupan de la información no estructurada. Sus capacidades incluyen:
  - Extracción de información y anotación de textos no estructurados (identificación de los eventos, personajes, nombres, lugares, marcas).
  - Clasificación automática y etiquetar los documentos.
  - Extracción automática y la construcción de un dominio específico de terminología léxicos.

  Con el motor Sophia se relaciona debido a que ambos tienen como objetivo la recuperación de la información y la anotación de texto sobre un documento, para esta propuesta también resulta de vital importancia este proceso.

  La diferencia la encontramos en como es el proceso de anotado, mientras que para este Software lo realiza de manera automática, el sistema que se desarrollará en esta propuesta permite al usuario seleccionar las etiquetas que se van a anotar sobre el documento.

- Asociación Mexicana de Procesamiento del Lenguaje Natural (AMPLN) [6].

  La AMPLN es una organización profesional no lucrativa, cuya misión es fomentar la interacción e intercambio de ideas entre especialistas mexicanos en el procesamiento del lenguaje natural, así como difundir los logros y la importancia entre la sociedad nacional. Esta misión es también uno de los objetivos que incorpora el desarrollo de este proyecto, donde se contempla contribuir aportando ideas y compartiendo las técnicas empleadas para la elaboración de este proyecto.

Descripción técnica

Se desarrollará un sistema de procesamiento de textos de investigación que están escritos en inglés usando técnicas de procesamiento del lenguaje natural. La Figura 1 muestra cómo se llevará a cabo este proceso y posteriormente describiremos cada módulo que compone al sistema.

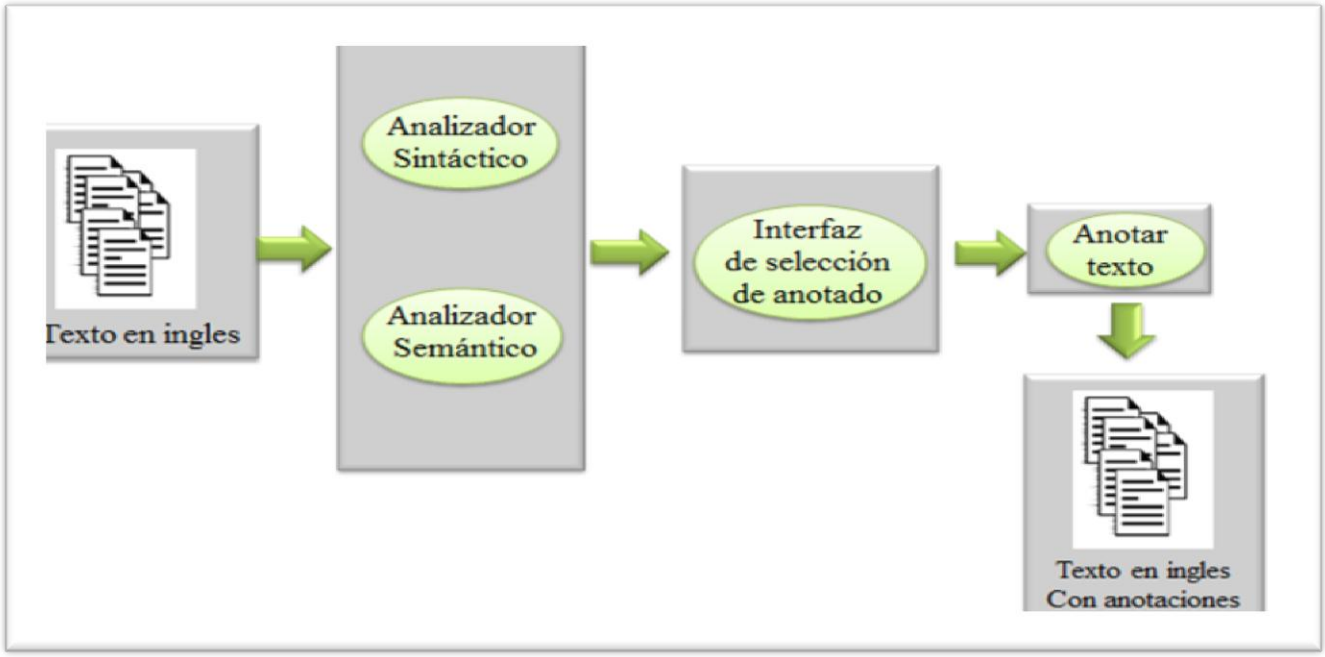

Figura 1: Esquema general del sistema de procesamiento de Textos

Para lograr que el objetivo del proyecto se cumpla y se realice de forma ordenada, dividimos todo el proceso en cuatro módulos, donde cada módulo tiene una función bien definida. El diagrama de la Figura 1 muestra el orden y la secuencia que tiene cada módulo. Cada módulo depende completamente del módulo anterior.

- Analizador Sintáctico

También conocido como Parser. Este módulo es el punto inicial de todo proceso, recibimos como parámetro al texto digitalizado y se revisa el correcto orden de la palabras y su afectación en el significado, este módulo obtendrá como resultado una lista de etiquetas sintácticas que principalmente estará conformada por: unidades léxicas, espacios, oraciones y elementos del enunciado como se puede observar en la Figura 2., en este módulo se divide todo el texto digitalizado,

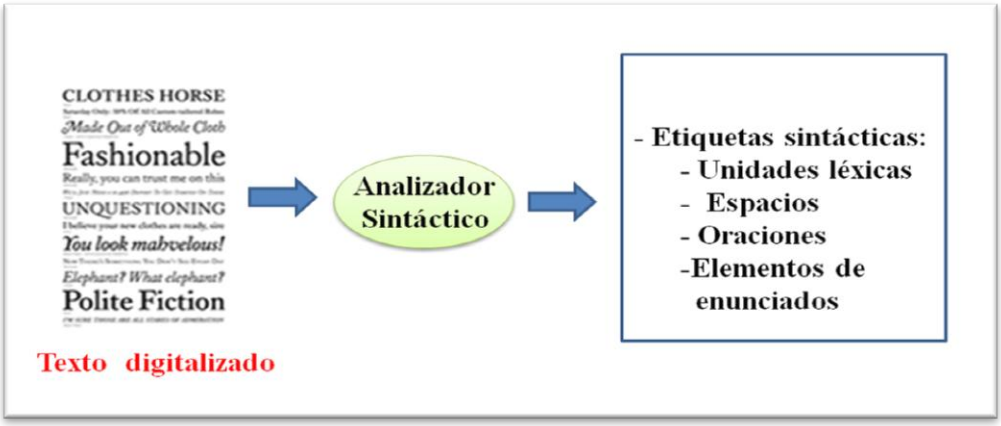

Figura 2: Parámetro de entrada y salida del módulo del analizador sintáctico.

En esta sección es donde realizaremos el análisis de frases del lenguaje natural contenidas en todo el documento, toda la información que se genere en este módulo servirá para el módulo del analizador semántico.

- Analizador Semántico

Como se observa en la Figura 3 en este proceso recibe como parámetro de entrada la lista de etiquetas sintácticas generadas por el analizador sintáctico. Teniendo esta lista se procede a estudiar el significado, sentido o interpretación literal de las palabras, frases y oraciones. Realizado este proceso, se seleccionará a aquellas palabras, frases y oraciones que son de interés para cumplir con el propósito del proyecto como lo son: Autor, Institución, Fecha de publicación, Palabras clave, correo electrónico, tipo de publicación (libro, tesis, reporte, artículo, etc.).

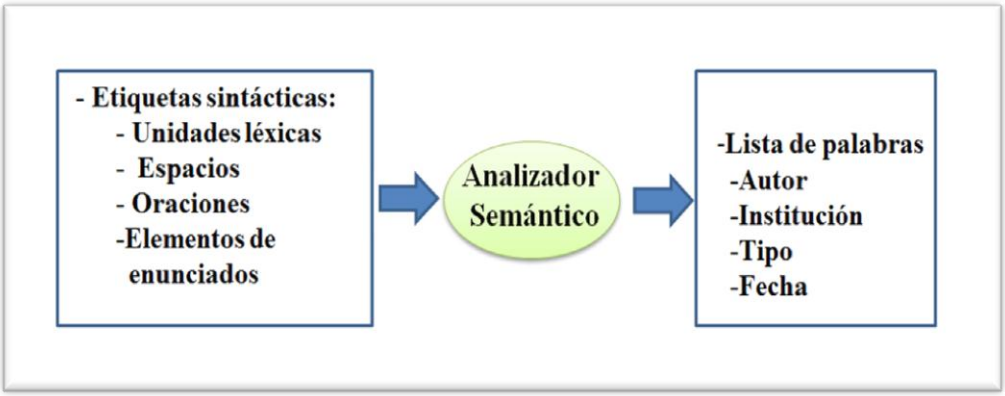

Figura 3: Parámetro de entrada y salida del módulo del analizador semántico.

Este módulo tendrá como salida una lista de palabras que se usarán como etiquetas para el archivo de salida.

- Interfaz de Selección de Anotado

Este módulo será el que permitirá al usuario interactuar con el sistema, en esta sección es donde convergen todos los procesos de todas las etapas que conforma este proyecto, En este módulo se le presenta al usuario de forma visual, la lista de etiquetas disponibles para ser almacenadas y que serán asociadas al mismo documento que el módulo anterior ha generado, otorgando al usuario la oportunidad de que el seleccione las etiquetas para ser anotadas, además esta interfaz permite seleccionar el documento que será procesado. Las funciones que principalmente realiza son:

- Seleccionar documento de entrada.
- Selección de etiquetas que serán anotados.

- Anotar Texto

Este es el último proceso del sistema, en esta etapa será donde, después de tener bien identificadas y seleccionadas las etiquetas, se realizará el anotado semántico utilizando estas etiquetas y que se almacenarán como información que están asociadas al mismo documento.

Especificación Técnica

Entre los aspectos de complejidad para el desarrollo del sistema se encuentran: el manejo indistinto de diferentes formatos de los documentos (PDF, DOCx, TXT, ODT, etc.), y el idioma en el que se encuentra el contenido del documento.

Otra limitante que tendrá es el tipo de etiquetas que puede agregar a los documentos, para cumplir el objetivo de este proyecto el sistema sólo contempla las siguientes: Autor, Institución, Fecha de publicación, Palabras clave, correo electrónico, tipo de publicación (libro, tesis, reporte, artículo, etc).

El desarrollo del sistema se escribirá en lenguaje java utilizando el entorno de desarrollo integrado (IDE por sus siglas en inglés) NetBeans IDE 7.3 Release, el cual es de software libre, además se usarán las herramientas de software libre: GATE y OpenNLP . Este sistema tendrá capacidad para procesar 500 documentos de máximo 100 páginas cada uno.

La librería de OpenNLP está basada en un conjunto de herramientas para el procesamiento de lenguaje natural, en este proyecto será utilizada para la separación en elementos independientes de las oraciones de texto contenidas en todo el documento. Otras de las herramientas que usaremos de esta librería serán las encargadas de detección de oraciones y anotado morfosintáctico que serán implementados en el módulo del analizador semántico.

GATE es un conjunto de herramientas para realizar tareas de procesamiento de lenguaje natural, esta herramienta permitirá facilitar la construcción de la lista de palabras del analizador sintáctico. El uso de esta herramienta facilitará la tarea de encontrar etiquetas sintácticas permitiendo la división en unidades léxicas del texto, identificar espacios, oraciones y elementos de enunciados.

Al concluir el proyecto terminal se entregarán tres discos compactos al Coordinador de Estudios de Ingeniería en Computación que incluirán el reporte final del proyecto terminal en un archivo PDF (sin restricciones) y el código fuente de la aplicación en un archivo comprimido (sin restricciones). El reporte final contendrá al menos: portada, resumen, tabla de contenido, objetivos, introducción, desarrollo del proyecto, conclusiones, bibliografía y apéndices. Los apéndices contendrán al menos un listado del código fuente desarrollado.

Calendario de Trabajo

En la Tabla 1 se describen las actividades que se realizarán para el "Trimestre 2013 Invierno", que corresponde a 9 horas de trabajo por semana, cumpliendo al finalizar el trimestre con un total de 99 horas.

Trimestre 13 Invierno	1	2	3	4	5	6	7	8	9	10	11	Horas
Recopilación de Documentos.												9
Instalación y configuración de herramientas a utilizar.												9
Diseño del analizador Sintáctico de textos												18
Implementación del analizador Sintáctico de Textos.												18
Diseño del analizador Semántico de textos												27
Implementación de del analizador Semántico de Textos.												9
Primera revisión												9

Tabla 1: Calendario de actividades para el Trimestre 2013 Invierno.

En la Tabla 2 se describen las actividades que se realizarán para el "Trimestre 2013 Primavera" se divide en espacios de 18 horas de trabajo por semana, cumpliendo un total de 198 horas de trabajo al finalizar el trimestre.

Trimestre 13 Primavera	1	2	3	4	5	6	7	8	9	10	11	Horas
Diseño de interfaz de usuario												36
Diseño de selección de anotado de texto												36
Implementación de interfaz de usuario												36
Implementación de anotando de texto												36
Segunda revisión y pruebas												18
Redacción del reporte final, manual de usuario y configuración												36

Tabla 2: Calendario de actividades para el Trimestre 2013 Primavera.

Recursos

Software con el que se cuenta para desarrollar este proyecto:

- NetBeans IDE 7.3 Release
- GATE
- OpenNLP

Hardware con el que se cuenta para desarrollar este proyecto:

- Se cuenta con una computadora personal con características:
- Sistema operativo: Windows 7 Professional, arquitectura 32 bits
- Procesador: Intel(R) Celeron (R) 1.73GHz
- Memoria RAM: 2GB

El asesor se responsabiliza de guiar al alumno y de que todos los recursos anteriormente citados estarán disponibles para el alumno, de modo que el proyecto terminal se pueda concluir en tiempo y forma

Asesor

Maricela Claudia Bravo Contreras
Profesor Asociado
Departamento de Sistemas

Bibliografía

[1] M. S. de J. Ugalde Chávez, "*Sistema de recuperación de información semántico*", Proyecto Terminal, División de CBI, Universidad Autónoma Metropolitana, Azcapotzalco, México, 2011.

[2] E. Sánchez Estrada, "*Clasificación de servicios web semánticos mediante ontologías"*, Proyecto Terminal, División de CBI, Universidad Autónoma Metropolitana, Azcapotzalco, México, 2012.

[3] S. Angeles Camacho, "*Sistema semántico para la búsqueda, selección y adecuación de contenidos educativos basados en perfiles de aprendizaje*", Proyecto Terminal, División de CBI, Universidad Autónoma Metropolitana, Azcapotzalco, México, 2011.

[4] B. Beatriz. (2007, Octubre 12). *Técnicas del Procesamiento del Lenguaje Natural* [en línea]. Disponible: http://cpti.azc.uam.mx/ProyectosTerminales/Propuestas/PropuestaSandra.pdf

[5] CELI. (2011). Sophia Semantic Engine. [En línea].
Disponible: http://www.celi.it/en/sophia-semantic-engine.shtml

[6] ACL. (2011). About the ACL. [En línea].
Disponible: http://www.aclweb.org/index.php?option=com_content&task=view&id=38&Itemid=35

[7] B. Coppin, "Understanding Language" in Artificial intelligence illuminated, Canada:
Jones and Bartlett Publishers, 2004