THE FRUIKIS

The brain never dream´s

Home

Procesamiento del Lenguaje Natural para recuperación de información

Introducción

En el último congreso internacional sobre Web Semántica, celebrado en Osaka, del 18 al 21 del mes pasado, la presencia de aplicaciones centradas en **Procesamiento de Lenguaje Natural** (PLN) fue más que notable. De hecho, Gate, una conocida aplicación para ingeniería **lingüística** diseñada en la universidad de Sheffield, fue una de las estrellas invitadas al figurar referenciada en un buen número de los trabajos presentados al congreso.

Ahora bien, la utilidad del **procesamiento natural del lenguaje** para la implementación de una Web Semántica, no es un descubrimiento de este año. En el pasado 2004, Ricardo Baeza-Yates firmo un interesante trabajo sobre la aplicación de técnicas de **procesamiento del lenguaje natural** para la **Recuperación de Información** donde proponía a la Web Semántica como una de las principales aplicaciones prácticas de técnicas convinadas de PLN(procesamiento del lenguaje natural) y RI(recuperación de información).

Muchos pueden argumentar que el propio padre de la idea defiende que la Web Semántica no es una Web basada en técnicas pertenecientes al área de **Inteligencia Artificial** (IA), pero esto no significa que no podamos utilizar estas técnicas como base y apoyo para la implementación de su idea, ya que, más allá de rencillas de carácter académico, todo lo que nos ayude a hacer realidad una nueva Web es útil independientemente de conceptualizaciones de carácter teórico.

Hay que tener en cuenta que, hoy por hoy la Web Semántica no existe como tal, más allá de implementaciones puntuales de carácter experimental. El hecho de que exista pasa inexorablemente por la generación de contenidos web semánticos que den cuerpo a la idea de una web más organizada. La generación de contenidos de carácter semántico no es asimilable de forma manual por lo usuarios y autores de la Web, por lo que es necesario la automatización de todas, o por lo menos parte de las tareas de generación de contenidos web semánticos. Es aquí donde el PLN y la RI tienen mucho que aportar, ya que permiten la implementación de aplicaciones capaces de generar información de tipo semántico que dote de cuerpo a la Web Semántica y la conviertan en una realidad.

Analizadores sintácticos, que permitan comprender la estructura de las frases de forma automática, etiquetadores léxicos, reconocedores de entidades como nombres, fechas lugares, todas ellas son herramientas automáticas esenciales para la generación de contenidos web semánticos, es más, me atrevo a decir que sin ellas no es posible una web semántica real, ya que el coste de elaboración manual de contenidos semánticos no es asimilable desde ningún punto de vista.

Procesadores de lenguaje natural

Como mencionamos anteriormente el Lenguaje Natural(LN) es el medio que utilizamos de manera cotidiana para establecer nuestra comunicación con las demás personas

Este tipo de lenguaje es el que nos permite el designar las cosas actuales y razonar a cerca de ellas, fue desarrollado y organizado a partir de la experiencia humana y puede ser utilizado para situaciones altamente complejas y razonar muy sutilmente. La riqueza de sus componentes semánticos da a los lenguajes naturales su gran poder expresivo y su valor como una herramienta para razonamiento sutil. Por otro lado la sintaxis de un LN puede ser modelada fácilmente por un lenguaje formal, similar a los utilizados en las matemáticas y la lógica. Otra propiedad de los lenguajes naturales es la polisemantica, es decir la posibilidad de que una palabra en una oración tenga diversos significados.

En un primer resumen, los lenguajes naturales se caracterizan por las siguientes propiedades:

1. Desarrollados por enriquecimiento progresivo antes de cualquier intento de formación de una teoría.
2. La importancia de su carácter expresivo debido grandemente a la riqueza del componente semántico(polisemantica).
3. Dificultad o imposibilidad de una formalización completa.

Las aplicaciones del Procesamiento de Lenguajes Naturales son muy variadas, ya que su alcance es muy grande, algunas de las aplicaciones del PLN son:

- **Traducción automática**: se refiere más que nada a la traducción correcta de un lenguaje a otro, tomando en cuenta lo que se quiere expresar en cada oración, y no solo palabra por palabra. Una aproximación a este tipo de traductores es el babylon.

- **Recuperación de la información**: en esta aplicación, un claro ejemplo seria el siguiente: Una persona llega a la computadora y le dice(en LN) que es lo que busca, esta busca y le dice que es lo que tiene referente al tema.

- **Extracción de Información y Resúmenes**: Los nuevos programas, deben tener la capacidad de crear un resumen de un documento basándose en los datos proporcionados, realizando un análisis detallado del contenido y no solo la truncando las primeras Lineas de los párrafos.

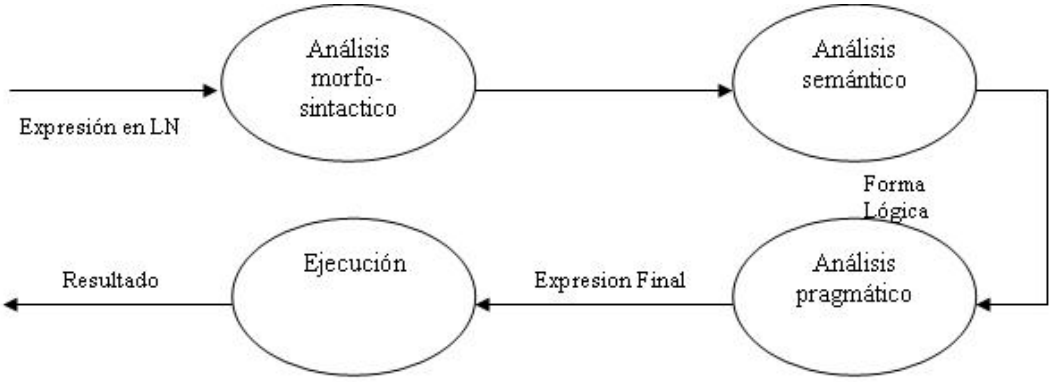
- **Resolución cooperativa de problemas**: La computadora debe tener la capacidad de cooperar con los humanos para la solución de problemas complejos, proporcionando datos e información, incluyendo también, la demanda de información por parte del ordenador al usuario, debiendo existir una excelente interactividad entre el usuario y el ordenador.

- **Tutores inteligentes**: La aplicación del PLN en este aspecto, vienen por computadora, debiendo esta ser aprox. en un 99%, al tener esta la capacidad de evaluar al educando y tener la capacidad de adaptándose a cada tipo de alumno.

- **Reconocimiento de Voz**: Esta es una aplicación del PLN que más éxito ha obtenido en la actualidad, ya que las computadoras de hoy ya tienen esta característica, el reconocimiento de voz puede tener dos posibles usos: para identificar al usuario o para procesar lo que el usuario dicte, existiendo ya programas comerciales, que son accesibles por la mayoría de los usuarios, ejemplo: ViaVoice.

Para continuar nuestro estudio de los lenguajes naturales, es necesario el que conozcamos los niveles del lenguaje, los cuales serán utilizados para la explicación de el siguiente tema que es la Arquitectura de un sistema de PLN. Los niveles de lenguaje que daremos a conocer son los siguientes: fonológico, morfológico, sintáctico, semántico, y pragmático.

- Nivel Fonológico: trata de cómo las palabras se relacionan con los sonidos que representan.
- Nivel Morfológico: trata de cómo las palabras se construyen a partir de unas unidades de significado mas pequeñas llamadas morfemas.
- Nivel Sintáctico: trata de cómo las palabras pueden unirse para formar oraciones, fijando el papel estructural que cada palabra juega en la oración y que sintagmas son parte de otros sintagmas.
- Nivel Semántico:
  trata del significado de las palabras y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independiente del contexto, es decir de la oración aislada.
- Nivel Pragmático: trata de cómo las oraciones se usan en distintas situaciones y de cómo el uso afecta al significado de las oraciones. Se suele reconocer un subnivel recursivo: discursivo, que trata de cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.



La explicación a la arquitectura mostrada para los sistemas PLN es sencilla:

1. El usuario le expresa a la computadora que es lo que desea hacer.
2. La computadora analiza las oraciones proporcionadas, en el sentid morfológico y sintáctico, es decir, si las frases contienen palabras compuestas por morfemas y si la estructura de las oracioneses correcta.
3. El siguiente paso, es analizar las oraciones semánticamente, es decir saber cual es el significado de cada oración, y asignar el significado de estas a expresiones lógicas.
4. Una vez realizado el paso anterior, ahora podemos hacer el análisis pragmático de la instrucción, es decir una vez analizadas las oraciones, ahora se analizan todas juntas, tomando en cuenta la situación de cada oración, analizando las oraciones anteriores, una vez realizado este paso, la computadora ya sabe que es lo que va a hacer, es decir, ya tiene la expresión final.
5. Una vez obtenida la expresión final, el siguiente paso es la ejecución de esta, para obtener así el Resultado y poder proporcionárselo al usuario.

Uno de los grandes problemas del PLN se produce cuando una expresión en lenguaje natural posee más de una interpretación, es decir, cuando en el lenguaje de destino se le pueden asignar dos o más expresiones distintas. Este problema de la ambigüedad se presenta en todos los niveles del lenguaje, sin excepción. Ejemplo:

"Juan vio a María, con el telescopio"

"Juan vio a María con el telescopio"

En apariencia este problema es demasiado sencillo, pero en realidad, es uno de los mas complicados y que más complicaciones ha dado para que el PLN pueda desarrollarse por completo, ya que al presentarse en todos los niveles del lenguaje, se tienen que desarrollar programas(en lenguaje formal) para solucionarlos en cada caso.

Entre las técnicas inductivas aplicadas para resolver estas tareas de desambiguación se puede encontrar el aprendizaje basado en ejemplos, aprendizaje basado en reglas de transformación, inferencia gramatical, y aproximaciones estadísticas basadas en modelos de máxima entropía o en modelos de Markov.

Estos últimos se han utilizado ampliamente en el campo del reconocimiento automático del habla tanto para el modelado acústico como para la construcción de modelos del lenguaje para el reconocimiento, tanto de palabras aisladas, como del discurso continuo. El éxito en estos sistemas y la disponibilidad de recursos ha permitido su extensión a los sistemas de PLN. Para poder llevar a cabo otras tareas de desambiguación en PLN utilizando modelos de Markov es necesario abordar cada una de éstas como problemas de etiquetado.

Además del etiquetado morfosintáctico, otros problemas como son el análisis sintáctico supercial o la desambiguación del sentido de las palabras, también pueden reducirse a un problema de etiquetado. Por

ejemplo, en la tarea de análisis supercial o chunking, el análisis de una oración puede representarse mediante etiquetas que indican a qué sintagma pertenece una palabra.

En este caso, la secuencia de observaciones pueden ser etiquetas morfosintácticas y los estados del modelo representan etiquetas de sintagma o de chunk. En caso de considerarse un análisis más complejo, como es el caso de la detección de cláusulas, pueden utilizarse etiquetas estructuradas que marquen el nivel de anidamiento de la palabra dentro del análisis. La desambiguación semántica puede verse como la asignación de la secuencia más probable de etiquetas semánticas (o sentidos) a las palabras de una oración.

Modelos ocultos de Markov

Los modelos ocultos de Markov
fueron desarrollados por A. Markov en 1913 para modelizar secuencias de palabras en ruso y en la actualidad se usan como herramienta estadística de propósito general.

Se formaliza la etiquetación como un proceso doblemente aleatorio parametrizable (los parámetros se pueden estimar de forma precisa en el entrenamiento) en el que el modelo del lenguaje es representado por un autómata finito probabilista.

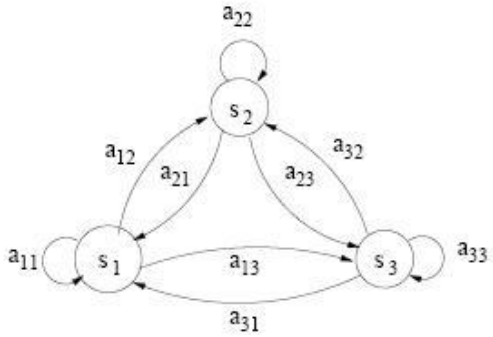


Figura 2.1: Modelo de Markov de 3 estados.

El modelo de comunicación este representado por la probabilidad de “emisión” de una palabra en un estado dado (la probabilidad de la palabra depende s´olo de la etiqueta), la descripción General del sistema se modeliza como un conjunto finito de estados, en el que pasado un intervalo de tiempo, el sistema cambia de estado de acuerdo a unas probabilidades asociadas a las transiciones entre estados.

Dos tipos de modelos:

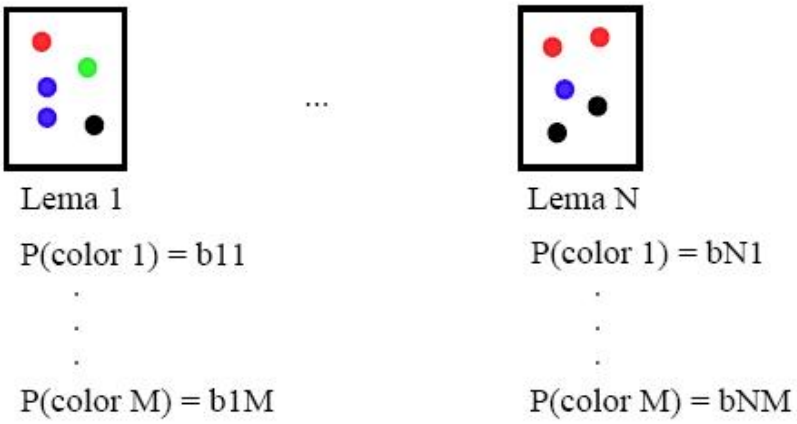
- Modelos Visibles:
  - Cada estado tiene asociado un ´unico proceso observable.
  - La salida del estado no es aleatoria.
- Modelos Ocultos:
  - En cada estado hay varios tipos de observaciones con diferentes probabilidades.
  - Modelo doblemente aleatorio:
    - a) trasiciones entre estados
    - b)observaciones asociadas.
  - Uno de los procesos no es observable directamente

Ejemplo:

Tenemos una serie de urnas en las que hay bolas de diferentes colores. No conocemos cuantas bolas de cada color hay en cada urna.

P(color 1) = b11
Lema 1
...

Lema N
P(color M) = b1M
...
P(color 1) = bN1
P(color M) = bNM
...



• Urnas = Estados
• Color = Observación

Distintas probabilidades de cada color en cada urna.Selección de urnas de acuerdo a las probabilidades de cada transición.Única salida observable: un conjunto de colores.Queremos saber cual es la secuencia de urnas más probable dada una secuencia de colores.

• Para modelizar etiquetas en PLN:

- Estados = Etiquetas (Urnas)
- Observaciones = Palabras (Colores)
- Secuencia de Observaciones = Frases del texto
- Instantes de tiempo = Posiciones dentro de la frase

Es posible una misma palabra (color) en distintas etiquetas (urnas), lo que da lugar a ambigüedades. El mismo color (palabra) puede aparecer más de una vez en cada urna, (etiqueta) dando lugar a distintas probabilidades de emisi´ón de palabras en cada etiqueta.

El algoritmo de Viterbi

El *algoritmo de Viterbi*
fue inicialmente desarrollado para encontrar, dada una secuencia de símbolos, la serie de transiciones más probable entre los estados de una cadena de Markov necesaria para producir dicha secuencia. Este problema es el equivalente markoviano al análisis sintáctico en una gramática regular estocástica.

El algoritmo de Viterbi es un caso particular del algoritmo de Programación Dinámica utilizado para encontrar un camino extremal en un grafo multietapa. Al igual que en el caso del análisis sintáctico para gramáticas regulares no deterministas, se recurre a un trellis, pero en este caso se define la función peso, no el dominio de los booleanos, sino en el intervalo [0..1], puesto que ahora representa la probabilidad de una regla o transición:

[rho]((j-1,u), (j,q)) [propersubset] [0..1]

y se sustituyen respectivamente las funciones "extremiza" por "max" y [circlemultiply] por el producto:

C(j,q) =

Al final del proceso C(n,|Q|) nos proporciona la probabilidad (de máxima verosimilitud) de que la cadena analizada pertenezca al lenguaje de la gramática.

Referencias interesantes:

http://gate.ac.uk/conferences/iswc2003/
http://gate.ac.uk/semweb.html
http://www.cc.gatech.edu/ccg/iswc05/
http://coleweb.dc.fi.udc.es/docencia/ln/
http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/markov/markov.pdf
http://www.dsic.upv.es/docs/bib-dig/tesis/etd-11262003-123346/TesisAntonioMolina.pdf

Descargas:

Procesamiento del Lenguaje Natural para recuperación de información.

Fecha ultima actualizacion: 18 de Abril de 2.007

Otros artículos de interes

- Evaluación de Buscadores Web
- Sistemas de Question-Answering
- Metadatos y documentos XML/RDF para recuperación
- Lenguajes de recuperación para la Web I
- Lenguajes de recuperación para la Web II
- Bases de datos nativas en Internet y sistemas para almacenar y recuperar documentos HTML, RDF y XML
- Modelos de recuperación I
- Motores de recuperación de documentos XML/RDF
- Usabilidad y Accesibilidad en el posicionamiento y en la recuperación de información, Extracción de información, Extracción y recuperación de información I
- Extracción y recuperación de información II
- Extracción y recuperación de información III
- Ontologías de metadatos y su fusión y mapeados de ontologías -> Minería de Textos

Autor: Alberto Martínez Mena ¿Tienes alguna duda, consulta o sugerencia ? Móndanos un email.