Procesamiento de Lenguajes Naturales

Introducción

La búsqueda de información sobre Internet se hace mediante el uso de motores de búsqueda. Dichos motores deben encontrar paginas web de interés para cada usuario. Los motores permiten al usuario de entrar palabras o frases para guiarle en su búsqueda de información. A partir de esas palabras o frases el motor debe reconstituir el entorno de la búsqueda. Por eso, necesitan una análisis de la entrada de cada usuario, un tal análisis requiere un conocimiento importante sobre la sintaxis, la semántica, la ortografía y la gramática del idioma. Entonces el procesamiento del lenguaje natural es una herramienta necesaria para cada motores de búsqueda sobre Internet.

El Procesamiento de Lenguajes Naturales (PLN) es una subdisciplina de la Informática y de la Lingüística que se carga de la aplicación de programas y técnicas informáticas a todos los aspectos de la comunicación entre humanos, o entre humanos y maquinas con lenguajes naturales. La aplicaciones relacionadas con este dominio son :

- Traducción automática
- Corrección de faltas ortografías
- Recuperación y extracción de la información de textos
- Comprensión del lenguaje
- Generación automatizada de textos
- Reconocimiento del habla
- Síntesis de voz
- Respuesta a preguntas

¿ Que es unau herramienta de Procesamiento de Lenguajes Naturales ?

Al principio, en los anos 50, el objetivo de los PLN estaba solamente la traducción automática de textos de un idioma a una otra. Pero, ahora su campo de utilización es mucho mas largo.

Debida a sus estructuras, a su caligrafía o al sentido de sus palabras, las idiomas tienen muchas ambigüedades. Cada frase o oración de un idioma tiene muchas informaciones implícitas que un motor de búsqueda debe descodar para cumplir su tarea. Las dificultades de comprensión son distintas en cada una de las idiomas. Las herramientas de Procesamiento del Lenguaje Natural abreviado PLN, en ingles Natural Language Processing (NPL) se cargan de quitar las ambigüedades y ayudar al motor de búsqueda comprender lo que quiere encontrar un usuario. Por eso, las técnicas utilizadas por las herramientas PLN, son varias.

Una herramienta de PLN, que utilizamos cada día seria un corrector ortográfico, como los de tratamiento de textos. Si una palabra no se encuentra en el diccionario, se miran las combinaciones de letras mas cerca de la palabra entrada anteriormente, y aparece una lista de elección que podrían convenir. A veces, tal correctores toman en cuenta la pronunciación fonética de la combinación de letras entrada para encontrar una palabra que se pronunciarían como la combinación de letras.

Quitar una ambigüedad sobre el sentido de una palabra se puede hacer mediante el uso de léxicos, o según el estudio del contexto de la palabra. Pero a veces no es suficiente, se requieren herramientas mas sofisticadas, que pueden exhibir la estructura de una frase. Tal herramientas estudian cada palabra de la frase para encontrar su función en la frase. Así, según la función de una palabra relativamente a una otra, pueden encontrar el sentido de la frase, y Así de la palabra. Aquellas herramientas son analizadores sintácticos y lematizadores (“taggers” en ingles).

Hay otras herramientas que esos, especialmente para idiomas con alfabetos distintos el chino, el ruso o el arabio, que tienen distintos problemas. Veremos un poco mas en detalle los analizadores sintácticos y lematizadores, que son las herramientas mas importantes de PLN para las idiomas del occidente utilizadas por los motores de búsqueda.

Analizadores sintácticos

El análisis de un texto consiste en exhibir la estructura de un texto, es decir dar informaciones mas o menos completas sobre la sintaxis de las frase de un texto.

Un analizador sintáctico es un programa informática que se carga de dicha tarea. Aquella operación supone una formalización del texto, que esta visto como un elemento de un lenguaje formal, et que esta regido por un conjunto de reglas componiendo una gramática formal. La estructura revelada por el análisis da precisamente como la reglas gramáticas están combinadas en el texto.

Para analizar un texto, se efectúan generalmente varias análisis para encontrar la mayor significación al texto. Así, se hace al principio una análisis léxico del texto, que le corta en elementos del léxico. Mas adelante, se realiza el análisis sintáctico encuentra las relaciones entre los diferentes elementos del léxico, identificados anteriormente con el análisis léxico. Al final, se hace un análisis semántica que estudia el sentido general del texto.

Software análisis sintáctico :

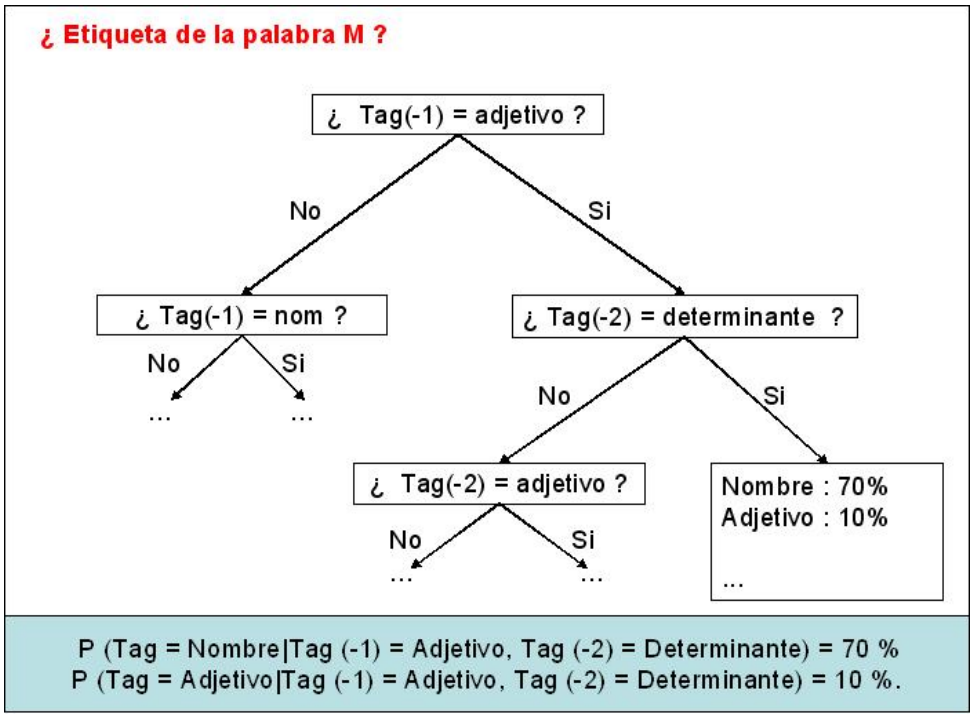
- Word mapper
- LEXICO 3
- Cordial
- alceste

lematizadores

Un lematizador es un software que permite, a partir de un texto etiquetado gramaticalmente, conocer el lema, es decir la raíz o la forma del diccionario de todas las palabras de un texto.

Las reglas para etiquetar un texto son relativas al etiquetador utilizado, pero generalmente, la estimación que una palabra tenga una etiqueta gramática se hace en relación con el contexto. A partir de una serie de tres etiquetas conocidas que constituyen el conjunto de aprendizaje, los etiquetadores utilizan arboles binarios construidos para saber cual es la clase mas probable para una palabra.

Ejemplo : Etiquetación por un lematizador



Las reglas para etiquetar las palabras pueden ser establecidas según el entrenamiento del etiquetador sobre un corpus de etiquetado a la mano. Es el caso de Winbrill que esta entrenado a partir del Wall Street Journal. Pero, en aquello caso el léxico puede ser no adaptado por textos especializados.

Cada etiquetador tiene ficheros de parámetros que utilizara para etiquetar mejor un texto. Aquellos ficheros de parámetros están generalmente compuestos de ficheros léxicos, reglas léxicas, reglas del contexto o excepciones en el idioma...

Una vez el texto etiquetado, se puede hacer la lematizacion. Entonces, se necesita encontrar la buena clase para cada palabra, sino una misma palabra podrá tener varias lemas.

Lematizadores :

- Tree Tagger
- Winbril
- Tnt Tagger
- Mbt Tagger