HISTORIA

Una de las ramas más importantes de la Inteligencia Artificial es aquella orientada a facilitar la comunicación hombre-computadora por medio del lenguaje humano, o lenguaje natural. El Procesamiento del Lenguaje Natural (PLN) es la disciplina encargada de producir sistemas informáticos que posibiliten dicha comunicación, por medio de la voz o del texto. Se trata de una disciplina tan antigua como el uso de las computadoras (años 50), de gran profundidad, y con aplicaciones tan importantes como la traducción automática o la búsqueda de información en Internet. Dado el tiempo disponible, es imperativo concentrar nuestros esfuerzos en un ámbito necesariamente limitado: los sistemas de PLN que utilizan técnicas de carácter estadístico aplicados al análisis del texto

Nació a finales de la década de los cuarenta, antes de que se acuñara la propia expresión «Inteligencia Artificial» (IA). No obstante, el PLN ha desempeñado múltiples papeles en el contexto de la IA, y su importancia dentro de este campo ha crecido y decrecido a consecuencia de cambios tecnológicos y científicos. Los primeros intentos de traducir textos por ordenador a finales de los cuarenta y durante los cincuenta, por ejemplo, fracasaron debido a la escasa potencia de los ordenadores y a la escasa sofisticación lingüística. Sin embargo, los esfuerzos realizados en las décadas de los sesenta y de los setenta para producir interfaces en lenguaje natural para bases de datos y otras aplicaciones informáticas obtuvieron un cierto grado significativo de éxito. La década de los ochenta y el principio de la de los noventa han visto resurgir la investigación en el terreno de la TA, investigación que ha conducido a sistemas susceptibles de ser explotados industrialmente. Estos progresos favorables se deben a una combinación de factores que van desde un enorme aumento en la potencia de los ordenadores en relación a su coste hasta modelos del lenguaje humano mejores y más susceptibles de ser tratados computacionalmente. Por otra parte, nunca ha sido mayor la necesidad de sistemas de PLN para procesar datos textuales, incluyendo traducción, clasificación, recuperación y extracción de información.

DESCRIPCION

El Procesamiento del Lenguaje Natural (PLN) es una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre/máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes formales. Todo sistema de PLN intenta simular un comportamiento lingüístico humano; para ello debe tomar conciencia tanto de las estructuras propias del lenguaje, como del conocimiento general acerca del universo de discurso. De esta forma, una persona que participa en un diálogo sabe cómo combinar las palabras para formar una oración, conoce los significados de las mismas, sabe cómo éstos afectan al significado global de la oración y posee un conocimiento del mundo en general que le permite participar en la conversación. En este curso se realiza una breve introducción al PLN presentando la organización de los sistemas de comprensión del lenguaje natural (módulos de análisis léxico, sintáctico y semántico) y las aplicaciones del PLN que coexisten actualmente en este campo (traducción automática, acceso a Bases

de Datos, extracción de información en Bases de Datos, recuperación o búsqueda de información, etc.).
Entre las tareas principales del procesamiento de lenguaje natural se puede mencionar:

TRADUCCION AUTOMATIZADA

A principios de la década de lo0s 60's habían bastantes esperanzas en el sentido de que las computadoras llegaran a ser capaces de traducir de un lenguaje a otro, de manera similar al proyecto de Tuning que era capaz de traducir mensajes codificados a un Alemán Claro. Sin embargo, ya para 1966 era evidente que para lograr una traducción es necesario contar con una comprensión del significado del mensaje (y por lo tanto, un minucioso conocimiento del mundo), en tanto que para una segmentación de código basta con conocer las propiedades sintácticas de los mensajes.
Una aplicación satisfactoria ha sido el sistema TAUM-METEO, diseñado por la universidad de Montreal y que traduce informes meteorológicos del ingles al Francés. La razón de su éxito es debido a que el lenguaje que se emplea en este tipo de informes meteorológicos gubernamentales se caracteriza por su irregularidad y estilo tan especifico.
En el caso de dominios más generales y amplios, los resultados obtenidos han sido menos espectaculares, un ejemplo seria el sistema SPANAM que logra realizar traducciones del ingles al español
Acceso a una Base de Datos
El primer logro obtenido por el PLN fue en el área de acceso a las bases de datos. Por 1970 las computadoras principales contaban con muchas bases de datos, pero su acceso se lograba solo escribiendo complicados programas en oscuros lenguajes de programación. El personal que estaba a cargo de de las unidades principales no tenia capacidad para responder a todas las solicitudes de los usuarios que requerían los datos, y desde luego que los usuarios no estaban dispuestos a aprender a programar sus solicitudes. Las interfaces para lenguajes naturales fue la solución al problema anterior
La primera de estas interfaces fue el sistema LUNAR, prototipo construido por William Woods y por su equipo de naves espaciales tripuladas de la NASA. Aunque el sistema nunca se utilizo en las condiciones reales para los que fue diseñado, en una prueba logro responder satisfactoriamente 78% de las preguntas que se le realizaron.
Las ventajas son obvias y las desventaja es que el usuario nunca podrá saber que frase de una consulta es el correcto y cual no es de la incumbencia del sistema.

RECUPERACION DE INFORMACION

La recuperación de información, la tarea consiste en escoger de entre un grupo de documentos aquellos que tengan relevancia en una consulta a veces el documento se representa por un reemplazo como es el titulo y una lista de palabras claves y/o un resumen. Actualmente son muchos los textos que aparecen en línea por lo que resulta más común utilizar la totalidad del texto, probablemente subdividir en secciones, cada una de las cuales sirve como documentos independientes para textos de recuperación. En los primeros sistemas de recuperación de información la consulta se realizaba como una combinación booleana de palabras reservadas.

Los sistemas modernos han cambiado del modelo booleano por un modelo de espacio vectorial, en el que década lista de palabras se maneja como un vector de un espacio N dimensional, donde N es la cantidad de especimenes del conjunto documental, en este modelo la consulta es simplemente una "Lingüística computacional del lenguaje natural", que podría manejarse como un vector cuyo valor es uno para estas cuatro palabras o términos como se les conoce en la RI y 0 para los demás términos. Por lo tanto , la localización de documentos es cuestión de comparar este vector respecto de un conjunto de otros vectores y dar cuenta de aquellos que se aproximan mas.

CATEGORIZACION DE TEXTOS

Las técnicas PLN han tenido éxito en una actividad relacionada con lo anterior; La clasificaciones de textos de acuerdo con determinadas categorías. Son diversos los servicios comerciales que de esta forma ofrecen el servicio de permitir el acceso a noticias transmitidas por cable.
La categorización de textos es compatibles con las técnicas del PLN en aquellos casos en donde no lo es la RI puesto que en las categorías son fijas, y , gracias a ello, los diseñadores del sistema pueden dedicar su tiempo a afinar el programa para un problema determinado.

OBTENCION DE DATOS DE UN TEXTO

El cometido de la obtención de datos consiste en tomar un texto en línea y deducir de el algunas aseveraciones que se puedan incorporar a una base de datos estructurada.
Si el PLN tiene tanto valor práctico, ¿a qué se debe que no dispongamos aún de PLN en todos los PC? En otras palabras, ¿por qué resulta difícil el PLN? La respuesta a esta pregunta es bastante compleja, pero una dificultad sobresale sobre las demás:
Dificultad principal: El lenguaje natural es localmente ambiguo, y la resolución de ambigüedades es necesaria para un procesamiento eficaz.
Consideremos, por ejemplo, la traducción al español de la palabra inglesa hit, para la cual existen múltiples traducciones posibles en función del contexto. He aquí tres de ellas:
1. He hit the nail with the hammer.=> «golpear» o «martillar» (Golpeó el clavo con el martillo).
2. The car swerved and hit the tree.=> «chocar» (El coche se desvió bruscamente y chocó contra el árbol).
3. The soldier fired and hit his target.=> «acertar» (El soldado hizo fuego y dio en el blanco).
¿Cómo podemos saber cuál significado de hit elegir en cada una de las frases anteriores? La respuesta obvia, «a partir del contexto oracional», no resulta suficientemente operativa para ser aplicada en un sistema de PLN. Debemos determinar exactamente cuál es el contexto oracional, qué conocimiento básico es necesario (como, por ejemplo, qué es un coche, o que el verbo swerve se refiere a un acontecimiento involuntario), y cómo ordenar esta información para poder decidir definitivamente sobre los significados de las palabras. Muchas investigaciones en el campo del PLN han estudiado métodos de resolver las ambigüedades léxicas mediante diccionarios, gramáticas, bases de conocimiento y correlaciones estadísticas. Además, la resolución de ambigüedades

léxicas no se limita a la traducción automática. La consulta a bases de datos y la recuperación documental también requieren la resolución de ambigüedades lingüísticas. Por ejemplo, si queremos documentos sobre cardiac arrest («paro cardíaco») deberemos generar —automáticamente— una consulta que busque también heart failure («fallo cardíaco») y otras expresiones sinonímicas o pseudo sinonímicas. No obstante, desearemos prescindir de significados espurios como arrests by police («detenciones policiales») que conducirían a la recuperación de cantidades abrumadoras de documentos irrelevantes para nuestros fines.
Aparte de la ambigüedad léxica, hay otros tipos de ambigüedades lingüísticas que resolver. Las más importantes son la ambigüedad referencial y la ambigüedad estructural. La primera tiene lugar cuando se utilizan pronombres o sintagmas nominales concisos para hacer referencia a objetos o eventos descritos previamente. En estos casos, el sistema de PLN debe determinar la entidad lingüística previa a que hacen referencia estas anáforas. La segunda es aún más frecuente; el caso más claro de este tipo de ambigüedad es el de la ambigüedad en el nivel de dependencia de los sintagmas preposicionales (PP-attachment). Consideremos las frases siguientes: remove the bolt with an Allen wrench («quitar el perno con una llave inglesa») y remove the box with blue lettering («quitar el recuadro con letras azules»). En el primer caso, el sintagma preposicional modifica al verbo como instrumento y, en el segundo, modifica al objeto directo como especificador. Se hace necesaria la semántica para distinguir ambas estructuras: llaves inglesas y letras desempeñan funciones semánticas diferentes.
A pesar de todo, la resolución de ambigüedades no es una tarea tan abrumadora como para imposibilitar el desarrollo de sistemas de PLN con fines prácticos. Se han construido ya sistemas de PLN para interfaces, procesamiento de texto y traducción, sobre todo para dominios claramente delimitados (especialmente dominios técnicos), para los cuales las relaciones semánticas pueden ser enumeradas.

COMPONENTES DE UN SISTEMA DE PROCESAMIENTO DEL LENGUAJE NATURAL.

En los sistemas reales de comprensión de texto, la entrada esta constituida por una secuencia de caracteres a partir de la cual se obtienen palabras. En la mayoría de los sistemas se aplica un procedimiento que consta de los siguientes pasos:

Caracterización, Análisis Morfológico, consulta de Diccionario y Corrección de Errores.

La Caracterización es un procedimiento mediante el cual la entrada se fragmenta en diversos elementos básicos: palabras y signos de puntuación.

Análisis morfológico: El análisis de las palabras para extraer raíces, rasgos flexivos, unidades léxicas compuestas y otros fenómenos.
El análisis morfológico es el procedimiento que consiste en describir una palabra en función de los prefijos, sufijos y raíces que están presentes en ella. Las palabras se generan de 3 maneras:

1-Morfología por Inflexión.- Refleja los cambios operados en una palabra necesarios para su adecuada inserción en un determinado contexto gramatical.

2-Morfología por Derivación.-Crea nuevas palabras a partir de otra que, por lo general, pertenece a otra categoría. Por ejemplo, el sustantivo "Cortedad" se deriva del adjetivo corto y del sufijo edad.

3-Composición.- Consiste en formar una nueva palabra mediante la unión de otras dos. Por ejemplo, limpiabotas es una palabra compuesta resultado de la unión de limpia y botas, a su vez, el término limpia se origina del verbo limpiar mediante la morfología de derivación.

LA CONSULTA DEL DICCIONARIO SE REALIZA POR CADA ELEMENTO BÁSICO CONSTITUTIVO.

La Corrección de Errores es realizada cuando no se localiza una palabra en el diccionario. Hay por lo menos 4 tipos de corrección de errores:

Para el primero se recurre a las reglas morfológicas mediante las que se conjetura la posible clase sintáctica a la que pertenece la palabra
En el Segundo, el uso de letras mayúsculas permite suponer que la palabra se trata de un nombre propio
En el tercero, se utiliza el conocimiento de que ciertos formatos especializados indican fechas, horas, números del seguro social, etc.
Por supuesto, ninguna arquitectura de PLN presenta un flujo de control que consista en una mera concatenación lineal de estos módulos funcionales. Para mayor eficiencia, los análisis sintáctico y semántico a menudo se entremezclan o cotejan mutuamente. En efecto, resulta más eficiente realizar llamadas a la semántica como rutina paralela, a fin de eliminar interpretaciones espurias, con lo cual se evita la generación de numerosos posibles análisis sintáctica o léxicamente ambiguos. Por otra parte, las diferentes tareas de PLN plantean requisitos diferentes en cuanto a su arquitectura. La consulta a bases de datos en LN, por ejemplo, utiliza normalmente los componentes analíticos, y devuelve el resultado de la consulta en forma de tabla sin generación de lenguaje. La traducción automática, en cambio, realiza el análisis lingüístico usando la gramática y diccionarios de una lengua, y la generación, mediante la gramática y diccionarios de una o varias lenguas diferentes. El paso que enlaza el análisis de la lengua de origen con la generación de la lengua de destino consiste bien en una representación semántica común (llamada a menudo interlingua),o bien en un proceso de transferencia (transfer) entre el resultado del análisis y el inicio de la generación.

ANÁLISIS GRAMATICAL EFICIENTE

Have the students in section 2 of computer science 101 take the exam.

(Aplique el examen a los estudiantes de la segunda sección de ciencias de la computación 101)
Have the students in section 2 of computer science 101 taken the exam. ?
(Ya presentaron el examen los estudiantes de la segunda sección de ciencias de la computación 101)

Aunque las primeras palabras son las mismas, su análisis gramatical es distinto ya que la primera es una orden y la segunda es una pregunta. Utilizando un algoritmo de análisis gramatical con avance de izquierda a derecha. El algoritmo trata de construir la estructura correcta de manera no determinista, habría que partir de la conjetura de que la primera palabra forma parte de una orden o bien de una pregunta; sabrá que su conjetura es correcta hasta llegar a la undécima palabra, "take/taken" (aplicar o presentar) si la conjetura del algoritmo fuese errónea será necesario que retroceda hasta la primera palabra. Aunque este tipo de reversión es inevitable, si nuestro algoritmo de análisis gramatical es eficiente evitara volver a analizar "los estudiantes de la segunda sección de ciencias de la computación 101" en cuanto *FN* cada vez que retrocede.

A nivel general, para mejorar la eficiencia se puede hacer lo siguiente:

1. No haga dos veces lo que puede hacer una vez
2. No haga una vez lo que pueda evitar hacer
3. No represente diferencias innecesarias

Una vez que sabemos que "los estudiantes de la segunda sección de ciencias de la computación 101" es una *FN* es conveniente dejar guardado dicho resultado en una estructura de datos conocida como **grafica.** Estos algoritmos se conocen como **analizadores de grafica** guardar los resultados en la grafica es una forma de programación dinámica que evita la duplicación del trabajo.

Veremos que el algoritmo de análisis de grafica se combinan los procesamientos en sentido descendente y ascendente.

El resultado de nuestro algoritmo es un bosque empacado.

PROCESAMIENTO COMPUTACIONAL DEL LENGUAJE NATURAL (PLN)

Una meta fundamental de la Inteligencia Artificial(IA), es la manipulación de lenguajes naturales usando herramientas de computación, en esta, los lenguajes de programación juegan un papel importante, ya que forman el enlace necesario entre los lenguajes naturales y su manipulación por una maquina.

La grafica es una estructura de datos para representar resultados parciales del proceso de análisis gramatical de manera que se les pueda volver a usar posteriormente. La grafica

de una oración formada por n palabras consta de n+1 vértices y varios bordes que unen los vértices entre si.

AMBIGÜEDAD

En este capitulo se amplio el rango de las construcciones sintácticas y las representaciones semánticas de que nos ocupamos. Esto nos permite un más amplio leguaje.

Las redes de creencia son la solución a uno de los problemas más arduos: como combinar las evidencias aportadas por fuentes diversas. Y también nos quedan pro resolver otros 2 problemas más:

1.- Decidir que evidencia se incorpora ala red.
2.- Decidir que hacer con las respuestas obtenidas.

La ambigüedad es intrínseca en las lenguajes naturales, tanto a nivel morfológico como sintáctico y semántico. En el caso de la sintaxis, el hecho de que una frase sea ambigua se traduce en que es posible asociar dos o más estructuras sintagmáticas correctas a dicha frase.

EJEMPLO:

Tomaremos una frase conocida: “Juan vio un hombre con un telescopio en una colina”. Diferentes ubicaciones de las subestructuras correspondientes a los fragmentos “con un telescopio” y “en una colina” llevan a diferentes estructuras sintagmáticas completas para la frase, todas ellas correctas.

EVIDENCIA SINTACTICA

Son modificadores tales como los adverbios y frases prepositivas dan lugar a considerable ambigüedad debido a que se les puede asociar a varias cabezas a la vez. EJEMPLO:

Lee asked Kim to tell Toby to leave on Saturday.

Lee le pido a Kim que le dijera a Toby que saliera el sábado.

EVIDENCIA LEXICA

Son muchas palabras ambiguas, pero no todos los sentidos de una palabra tienen una misma posibilidad. EJEMPLO:
Si se pregunta cual es el significado de la palabra en ingles “PEN” la mayoría responderá que es un instrumento de escritura.

EVIDENCIA SEMANTICA

Es la probabilidad a priori del sentido de una palabra normalmente es menos importante que la respectiva probabilidad condicional en un contexto determinado.

EJEMPLO:

ORACION	**RELACION**
Comí espagueti con albóndigas	(ingrediente del espagueti)
Comí espagueti con ensalada	(plato para acompañar el espagueti)
Comí espagueti con desenfreno	(manera de comer)
Comí espagueti con un tenedor	(instrumento para comer)
Comí espagueti con un amigo	(acompañante)

Metonimia

Es usar un objeto para representar otro.

Ejemplo:
"Chrysler presento un nuevo modelo"

METAFORA

Es una figura retórica en al cual se emplea una frase con un determinado sentido literal para dar entender otro por medio de una analogía.

ARQUITECTURA DE UN SISTEMA DE PROCESAMIENTO DEL LENGUAJE NATURAL

Uno de los elementos fundamentales en el diseño de un sistema PLN es sin lugar a dudas la determinación de la arquitectura del sistema, es decir, como se introducen los datos a la computadora y como ella interpreta y analiza las oraciones que le sean proporcionadas.

a. El usuario le expresa (de alguna forma) a la computadora que tipo de procesamiento desea hacer;
b. La computadora analiza las oraciones proporcionadas, en el sentido morfológico y sintáctico;
c. Luego, se analizan las oraciones semánticamente, es decir se determina el significado de cada oración;
d. Se realiza el análisis pragmático del texto. Así, se obtiene una expresión final.

   Se ejecuta la expresión final y se entrega al usuario para su consideración.

COMPRENSION DEL DISCURSO

¿Que es un discurso?

En un sentido técnico, un discurso o un texto es una cadena del lenguaje, por lo general con extensión superior a una oración.

- Novelas
- Informes
- Conversaciones

Hasta ahora se ha ignorado en buena medida los problemas del discurso, prefiriendo la disección del lenguaje en oraciones individuales.

Pasos de el hablante para producir un discurso:

Intención -> generación-> síntesis.

Pasos del escucha dentro del discurso:

Percepción->análisis->desambiguación->desincorporacion.

El estado de conocimiento del escucha juega un papel crucial en la comprensión: dos agentes con distintos tipos de conocimiento entenderán de manera distinta el mismo texto.

Existen 6 tipos de conocimiento para poder lograr la comprensión:

1. Conocimiento general del mundo.
2. Conocimiento general sobre la estructura del discurso coherente.
3. Conocimiento general sobre la sintaxis y la semántica.
4. Conocimiento especifico sobre la situación de que este hablándose.
5. Conocimiento especifico sobre las creencias de los personajes.
6. Conocimiento especifico sobre las creencias del hablante.

LA ESTRUCTURA DEL DISCURSO COHERENTE.

En lógica, la conjunción es conmutativa, por lo que no hay diferencia entre P^R^Q y R^Q^P. Sin embargo, lo anterior no es valido en el caso de los lenguajes naturales.

Necesitamos una teoría sobre como armar discursos. Diremos que los discursos están formados por segmentos, pudiendo ser estos cláusulas, una oración completa o un grupo de varias oraciones consecutivas. Varias teorías están basadas en la noción de que cada segmento tiene que ver con el anterior mediante una relación de coherencia que define el papel de cada segmento.
Teoría de Hobbs(1990).

- El hablante desea transmitir un mensaje.
- Para hacerlo, el hablante tiene una motivación o meta.
- El hablante desea facilitar al oyente la comprensión del mensaje.
- El hablante debe vincular la información nueva con lo que el escucha ya sabe.

La teoría de Grosz y Sidner(1986) también explica en donde se enfoca la atención del hablante y del oyente a lo largo del discurso.

GENERACIÓN DE TEXTOS

El complemento natural a la capacidad de entender el lenguaje es el segundo componente de la comunicación, que es la capacidad de producir el texto o bien el habla. En cierto grado es una tarea más simple que la comprensión, ya que por lo menos la computadora puede elegir las expresiones que sabe producir.

Uno podría pensar que para la generación de texto sólo es suficiente saber las reglas de gramática, es decir, saber palabras de cuales números, tiempos y géneros hay que usar en la oración y en que orden ponerlas.
Sin embargo, hay algunos problemas en la generación de texto. Uno reside en la necesidad de elegir las palabras y expresiones que «se usan» en el contexto dado. El otro problema es que el texto producido con los métodos de fuerza bruta es aburrido, incoherente y a veces no entendible.

El propósito del lenguaje es transferir conocimientos de una persona a otra. El conocimiento es una estructura compleja, multidimensional, que usualmente se representa como una red, o grafo, de conceptos. Pero el modo que usamos para transferir el conocimiento es unidimensional: en cada momento sólo podemos decir un sonido, una letra. Entonces, el trabajo del lenguaje es codificar el conocimiento multidimensional en una cadena de letras, y después, en el cerebro del escuchante o el lector, decodificar esta secuencia en el conocimiento original.

El lenguaje es una estructura muy compleja. Afortunadamente, el codificador y decodificador funcionan en pasos, construyendo las estructuras más complejas de bloques más simples:

Palabras de letras,– Oraciones de palabras,–Textos de oraciones.–

APLICACIONES DEL PLN

- 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 truncando las primeras líneas 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, viene siendo más académico, ya que se refiere a la enseñanza asistida 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.