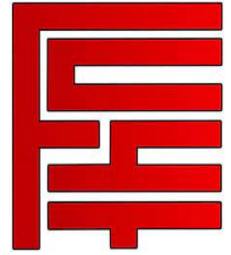


Universidad Mayor de San Simón
Facultad de Ciencias y Tecnología



Procesamiento de Lenguaje Natural

Estudiante: Anzaldo Alvarado Rodrigo
Chamaca Olivar Cleysi Maná
Coria Rojas Joseph David

Docente: Lic. Carmen Rosa Garcia Perez

Materia: Inteligencia Artificial II

Cochabamba-Bolivia

1. Introducción

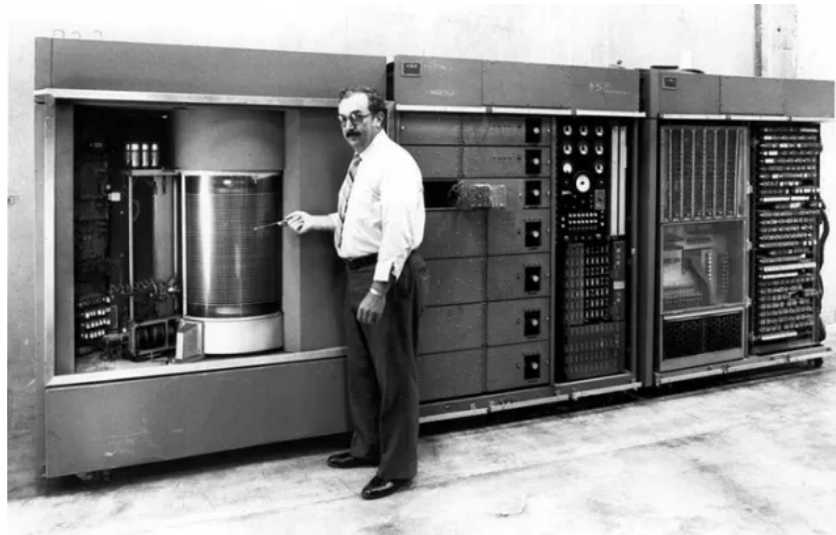
El campo del procesamiento del lenguaje natural (PLN) es un área muy extensa en la que están involucradas disciplinas de diversa naturaleza. Lo que distingue al procesamiento de lenguaje de otras áreas es el uso de conocimiento del lenguaje: El reconocimiento y síntesis de voz requiere de conocimiento acerca de la fonética y fonología de cómo se pronuncian las palabras, el reconocer las variantes que tienen las palabras requiere de conocimientos de morfología, se requiere de un conocimiento de la estructura del lenguaje y cómo se agrupan las palabras o su sintaxis, para entender el significado de cada palabra se requiere de conocimiento del significado léxico y de un conjunto de palabras de semántica composicional.

El PLN se encuadra dentro de la clase de problemas denominados «IA duros» (análoga a la clase de problemas NP duros). Es decir, el desarrollo de un sistema capaz de tratar el lenguaje en la misma forma en que lo haría un humano, pasa por la resolución del «problema de la IA». El mayor reto con el que se enfrentaría sería la resolución de las ambigüedades pragmáticas, ya que éstas requieren de complejos procesos de razonamiento sobre grandes bases de conocimiento de sentido común. Por supuesto, existen otros retos, como son la construcción de diccionarios, gramáticas y sistemas de análisis adecuados para el PLN.

2. Historia del Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN) es una disciplina con una larga historia. Nació en los años 50 como una subárea de la Inteligencia Artificial y la Lingüística, con el objetivo de estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural. Aunque se pueden encontrar trabajos de épocas anteriores, fue en 1950 cuando Alan Turing publicó un artículo titulado "Intelligence" en el que proponía lo que hoy se llama el test de Turing como criterio de inteligencia.

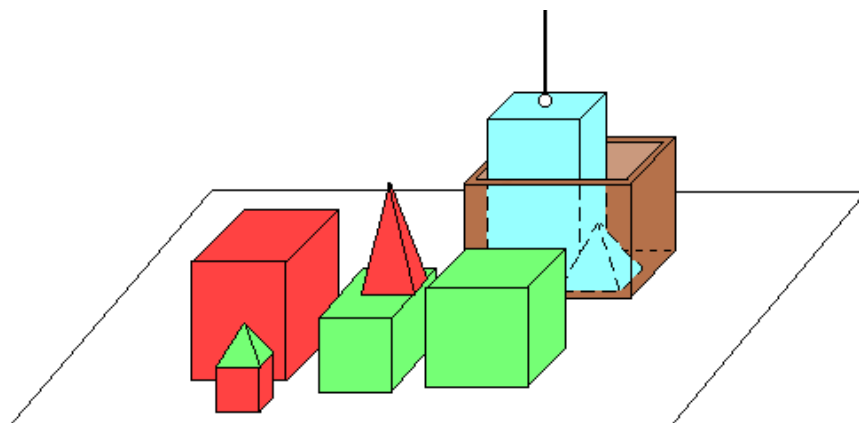
La PNL es una técnica utilizada para salvar la brecha de comunicación entre el ordenador y el ser humano y tiene sus orígenes en las primeras ideas de traducción automática (TA) que nacieron durante la Segunda Guerra Mundial. El experimento de Georgetown de 1954 consistió en la traducción automática de más de sesenta frases rusas al inglés en una colaboración entre IBM y la Universidad de Georgetown.



El Experimento Georgetown-IBM fue la primera máquina de traducción automática

Los autores afirmaban que en tres o cinco años la traducción automática sería un problema resuelto. Sin embargo, los avances reales fueron mucho más lentos y, tras el informe ALPAC de 1966, que constató que los diez años de investigación no habían estado a la altura de las expectativas, la financiación de la traducción automática se redujo drásticamente.

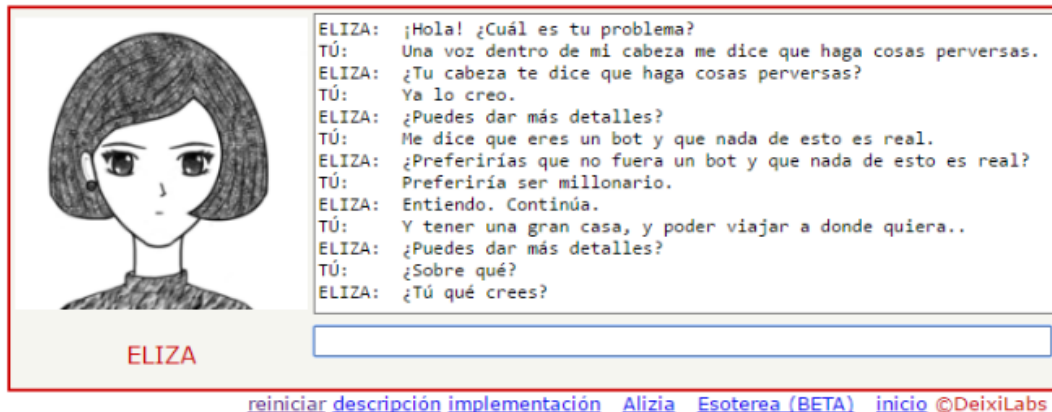
Entre los sistemas de procesamiento del lenguaje natural desarrollados en la década de los 60 destaca SHRDLU. SHRDLU era un sistema de lenguaje natural que trabajaba con "bloques" en un marco de vocabulario restringido. Uno podía darle instrucciones en lenguaje coloquial como "¿Puedes poner el cono rojo encima del bloque verde?" y el programa entendía que debía ejecutar la acción.



Otro ejemplo fue el programa ELIZA, una simulación del psicoterapeuta Carl Rogers, escrita por Joseph Weizenbaum entre 1964 y 1966. Sin información sobre el pensamiento o las emociones humanas, ELIZA proporcionaba una interacción sorprendentemente humana. Cuando el "paciente" superaba la pequeñísima base de conocimientos, ELIZA podía dar

una respuesta genérica, por ejemplo, responder a "Me duele la cabeza" con "¿Por qué dice que le duele la cabeza?".

ELIZA



En los años 70, muchos programadores empezaron a escribir "sistemas conceptuales", que estructuran la información del mundo real en datos comprensibles para el ordenador, y en los 80 la mayoría de los sistemas de procesamiento del lenguaje natural se basaban en complejos conjuntos de reglas escritas a mano. Sin embargo, debido al aumento constante de la potencia de cálculo y a la disminución gradual del predominio de las teorías chomskianas (lenguaje estructural), a finales de la década de 1980 se produjo una revolución en el procesamiento del lenguaje natural con la introducción de algoritmos conocidos como algoritmos de aprendizaje automático. Había nacido un nuevo paradigma. En los años 90, la popularidad de los modelos estadísticos aumentó de forma espectacular. La investigación reciente se ha centrado cada vez más en los algoritmos de aprendizaje no supervisado y semi supervisado. Estos algoritmos pueden aprender a partir de datos que no han sido anotados a mano con las respuestas deseadas, o utilizando una combinación de datos anotados y no anotados. En general, esta tarea es mucho más difícil que el aprendizaje supervisado, y suele producir resultados menos precisos para una cantidad determinada de datos de entrada.

3. ¿Que es el Procesamiento del Lenguaje Natural?

El Procesamiento del Lenguaje Natural es el campo de conocimiento de la Inteligencia Artificial que se ocupa de investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino, debiendo ésta entender las oraciones que le sean proporcionadas, el uso de estos lenguajes naturales, facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje. Virtualmente, cualquier lengua humana puede ser tratada por los ordenadores.

Lógicamente, limitaciones de interés económico o práctico hace que solo las lenguas más habladas o utilizadas en el mundo digital tengan aplicaciones en uso.

El procesamiento de lenguaje natural es fundamental para analizar a profundidad los datos de texto y voz de manera eficiente. Puede resolver las diferencias en dialectos, jerga e irregularidades gramaticales típicas en las conversaciones cotidianas. Las empresas lo utilizan para varias tareas automatizadas.

Según el documento EL PROCESAMIENTO DEL LENGUAJE NATURAL APLICADO AL ANÁLISIS DEL CONTENIDO DE LOS DOCUMENTOS, el procesamiento del lenguaje natural consiste en el estudio y análisis de los aspectos lingüísticos de un texto a través de programas informáticos. Un sencillo ejemplo de PLN es un corrector ortográfico de un procesador de textos que todos hemos empleado alguna vez, aunque hay otras herramientas más complicadas. Menciona Verdejo (1994, p.5) que el lenguaje natural se distingue de los lenguajes artificiales por su riqueza (en vocabulario y construcciones), flexibilidad (reglas con múltiples excepciones), ambigüedad (pudiendo darse diversos significados de una palabra o una frase según el contexto), indeterminación (permitiendo referencias y elipsis) y posibles interpretaciones del sentido literal según la situación en que se produce. Lo que son ventajas para la comunicación humana se convierten en problemas a la hora de un tratamiento computacional, ya que implican conocimiento y procesos de razonamiento que aún no sabemos ni cómo caracterizarlos ni cómo formalizarlos.

3.1. Funcionamiento del Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN) hace uso de la Inteligencia Artificial y consiste en transformar el lenguaje natural en un lenguaje formal, como el de la programación, que los ordenadores puedan procesar.

Normalmente, parte de dividirlo en elementos (frases, palabras, etc.) e intentar entender las relaciones entre ellos. Normalmente los humanos manejamos datos no estructurados en todos los sectores: artículos, informes, comentarios, mensajes o contratos. Y todos estos textos “escondan” información que puede ser útil para el negocio.

El Lenguaje Humano está repleto de ambigüedades que resulta difícil escribir un software que determine con precisión el significado deseado de los datos de texto a voz.

El PLN funciona a través del aprendizaje automático (ML o machine learning). Los sistemas de aprendizaje automático almacenan las palabras y las formas en que se unen como cualquier otra forma de datos. Frases, oraciones y a veces libros enteros se introducen en los motores de ML donde se procesan en base a reglas gramaticales, los hábitos lingüísticos de la vida real de la gente, o ambos. El ordenador utiliza estos datos para encontrar patrones y extrapolar lo que viene después. Tomemos el software de traducción, por ejemplo: En francés, "Voy al parque" es "Je vais au parc", así que el aprendizaje

automático predice que "Voy a la tienda" también comenzará con "Je vais au". Todo lo que la computadora necesita después de eso es la palabra para "tienda".

El PLN permite a las computadoras entender el lenguaje natural como lo hacen los humanos. Tanto si el lenguaje es hablado como escrito.

El PLN consta de dos fases principales: El preprocesamiento de datos y el desarrollo de respuestas por medio de algoritmos.

Preprocesamiento de datos: Consiste en preparar y limpiar los datos de texto para que las máquinas puedan analizarlos. El preprocesamiento pone los datos en forma viable y destaca las características del texto con las que se puede trabajar un algoritmo, Hay varias formas de hacerlo entre ellas:

- **Tokenización:** Se trata de dividir el texto en unidades mas pequeñas con las que se puede trabajar.
- **Eliminación de las palabras de parada:** En este caso se eliminan las palabras comunes del texto para que se queden las palabras únicas que ofrecen información sobre el texto
- **Lematización y Stemming:** En este caso, las palabras se reducen a su raíz para poder ser procesadas
- **Etiquetado de parte del discursos:** Es cuando las palabras se marcan en función de la parte del discurso que son, como sustantivos, verbos y adjetivos.

Una vez recolectada la información para procesarla hay muchos sistemas, los principales son:

- **Sistema Basado en Reglas:** Trabajan mediante la aplicación de reglas, comparación de resultados y aplicación de las nuevas reglas basadas en situación modificada. También pueden trabajar por inferencia lógica dirigida, bien empezando con una evidencia inicial en una determinada situación y dirigiéndose hacia la obtención de una solución, o bien con hipótesis sobre las posibles soluciones y volviendo hacia atrás para encontrar una evidencia existente.
- **Sistemas basados en Machine Learning:** Que hace análisis de datos y automatiza la construcción de modelos analíticos. Es una rama de la inteligencia artificial basada en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con mínima intervención humana.

4. Modelos PNL

Modelo.- Es un sistema que utiliza algoritmos y técnicas de aprendizaje para manejar el lenguaje de forma computacional, cada modelo puede realizar diversas tareas como la traducción, extracción de datos, etc.

Modelo basado en reglas.- Son modelos que tienen un conjunto de reglas, es decir una sintaxis fuertemente definida para el manejo de LN, son costosos y requieren gran tiempo y esfuerzo de mantenimiento más su alcance y manejo de lenguaje natural complejo es limitado. Estas reglas deben ser creadas por expertos en el tema de lingüística un ejemplo sería que estos modelos pueden identificar los nombre propios o verbos de un textos, mediante las reglas impuestas en el caso de nombres esser que comience con una letra mayúsculas o en el tema de los verbos que tenga terminaciones en -ar -er -ir.

Modelo estadístico.- Aca se usan técnicas de machine learning para recoger conjuntos de datos y generar patrones estadísticos de LN, estos modelos utilizan técnicas de análisis de texto, minería de texto y estadística para identificar patrones en el LN.

Estos modelos se entrenan en grandes conjuntos de datos en lenguaje natural y utiliza la estadísticas para determinar la probabilidad de una secuencia de palabras, esto cae en la predicción de oraciones, las aplicaciones de este modelo parten desde el reconocimiento de voz, corrección de ortografía y traducción automática.

El análisis de sentimientos es donde se determina el sentimiento de un texto, esto es etiquetando patrones o pequeños textos, sus aplicaciones van desde el monitoreo de redes sociales, análisis de opiniones de clientes o productos.

Sus dificultades son la comprensión del contexto de pequeños matices del lenguaje natural.

Modelo basado en redes neuronales.- Son modelos que basan su arquitectura en redes neuronales profundas que permiten un aprendizaje jerárquico y una representación más compleja del lenguaje natural.

Estos modelos se basan en algoritmos de deep learning, como redes neuronales convolucionales o redes neuronales recurrentes.

Al igual que los modelos estadísticos se busca predecir la siguiente palabra en una oración pero su ventaja es su mayor precisión en hacerlo debido al aprendizaje de

patrones, también existe la traducción neuronal donde se busca la traducción de un lenguaje a otro.

Actualmente el modelo de red neuronal más famoso es el de la generación de texto, donde se utiliza una red neuronal degenerativa para crear texto original, el texto generado es coherente y relevante sin caer en redundancia, sus aplicaciones van en la redacción automática de noticias etc.

En el análisis de sentimientos se utilizan las redes neuronales para clasificar un tono emocional del texto y clasificarlo como positivo, negativo o neutro este modelo es más preciso que el estadístico debido a que tiene una mayor capacidad de capturar relaciones más complejas en el lenguaje, a diferencia de los modelos estadísticos que se basan en características simples, las redes neuronales analizan el contexto y los patrones mas complejos, tambien la capacidad de generalizar nuevos datos que pueden contener “nuevas emociones”.

Modelos de aprendizaje por refuerzo.-Estos modelos se basan en la retroalimentación que reciben del entorno mejorando su rendimiento para tareas o temas del lenguaje natural específicos, este enfoque fue usado para la creación de chatbots .

Modelo GPT.- Modelo basado en arquitectura de transformadores, la arquitectura de transformadores es un tipo de red neuronal de aprendizaje profundo que utiliza múltiples capas de atención. La atención es un mecanismo que permite que la red se centre en las partes más específicas de una entrada de texto, para hacer predicciones más precisas. Las capas de atención se organizan en bloques de codificadores y decodificadores que se apilan juntos para formar la arquitectura del modelo.

El modelo usa un pre-entrenamiento basado en el aprendizaje sin supervisión, significa que entrena los datos sin etiquetas previas, dado un texto sin procesar este modelo se entrena para predecir la siguiente palabra en la oración de un párrafo, así genera texto coherente.

GPT tiene un montón de parámetros con valores ajustables para hacer las predicciones, algunos de estos parámetros son los embedding, capas de atención, capas de transformación, funciones de activación y regularización.

5. Componentes PNL

En PLN existen varios componentes o análisis diferentes, pero esto no significa que será necesario todos para el desarrollo, se usará dependiendo de los objetivos de la aplicación.

Análisis morfológico o léxico. Consiste en describir la forma, la composición y el significado de una palabra, es decir, el análisis interno de las palabras que forman oraciones para extraer lemas, rasgos flexivos, unidades léxicas compuestas. Es esencial para la información básica: categoría sintáctica y significado léxico.

Análisis sintáctico. Son reglas del lenguaje que permiten conocer y analizar cual es la estructura de la oración.

Análisis semántico. Es la que proporciona el significado de las oraciones, basándose en el sentido de las palabras que lo componen y una vez eliminadas las ambigüedades morfosintácticas.

Análisis pragmático. Se analizan diferentes variables relevantes para la comprensión de un enunciado, incluyendo el análisis del contexto de uso a la interpretación final.

Funcionamiento.

Para un sistema conversor de texto a voz no es necesario aplicar todos los componentes de PLN no es necesario un análisis semántico o pragmático, pero para un sistema conversacional requiere información muy detallada del contexto por lo tanto , es necesario aplicar todos los componentes de PLN. Una forma corta de explicar el funcionamiento de este sistema, es el siguiente:

1. El usuario expresa a la computadora lo que desea hacer.
2. En este paso se analizan las oraciones en el sentido morfológico y sintáctico, el análisis morfológico verifica si las frases contienen palabras compuestas por morfemas y el análisis sintáctico se encarga de verificar que la estructura de las oraciones sea la correcta. En esta etapa se utilizan tanto el analizador lexicográfico como el analizador sintáctico, donde el analizador lexicográfico primeramente identifica los componente léxicos definidos a priori, este analizador se denomina scanner, y el analizador sintáctico denominado parser verifica que se cumpla un cierto orden gramatical ya definido entre los elementos que identificó el scanner.
3. A continuación se procede al análisis semántico, en el cual a las oraciones ya verificadas en el paso anterior se les proporcionará un significado y posterior asignar los significados a expresiones lógicas.
4. Como siguiente paso realizamos el análisis pragmático, una vez analizadas las oraciones, ahora se analizan todas las juntas, tomando en cuenta la situacion de cada oracion, analizando las oraciones anteriores, es decir, se analizan las oración con su contexto correspondiente, de esta manera no solo tendríamos el significado de oraciones independientes, sino el significado de estas oraciones involucrando un contexto anterior.

5. Una vez obtenida la expresión final, el siguiente paso es la ejecución de esta, para obtener así el resultado y poder proporcionarle al usuario.

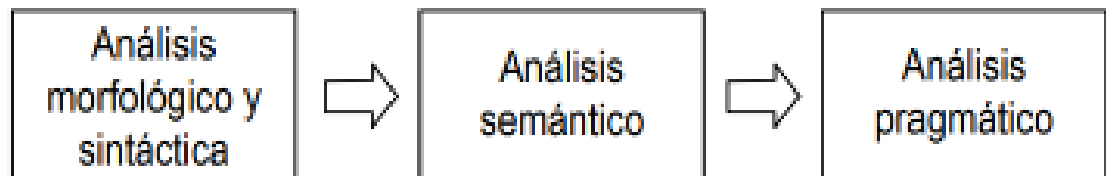


Figura N.º 3. Arquitectura de un Sistema de Procesamiento de Lenguaje Natural

Conceptos.

Lexema.- Unidad con significado léxico, aporta a la palabra una idea comprensible para los hablantes, expresa el significado común a toda la familia y puede coincidir o no con una palabra entera.

Ejemplo:

ZAPATo, ZAPATería, ZAPATero

Morfemas.- Unidad más pequeña de la lengua que tiene significado léxico o gramatical y no puede dividirse en unidades significativas menores, complementa al lexema en género, número, aumentativo, diminutivo y otras terminaciones.

Nub =(Raíz-Lexema)

Nube = (Morfema)

Nublado = (Morfema)

Nubosidad = (Morfema)

Analizador lexicográfico.- Se encarga de buscar componentes léxicos o palabras que componen el programa fuente, según unas reglas o patrones. La entrada del analizador léxico podemos definirla como una secuencia de caracteres, y produce una salida compuesta de tokens (componentes léxicos) o símbolos. Estos tokens sirven para una posterior etapa del proceso de traducción, siendo la entrada para el analizador sintáctico (parser).



gramática (N, T, P, S)

N ⇔ Símbolos no terminales.

T ⇔ Símbolos terminales

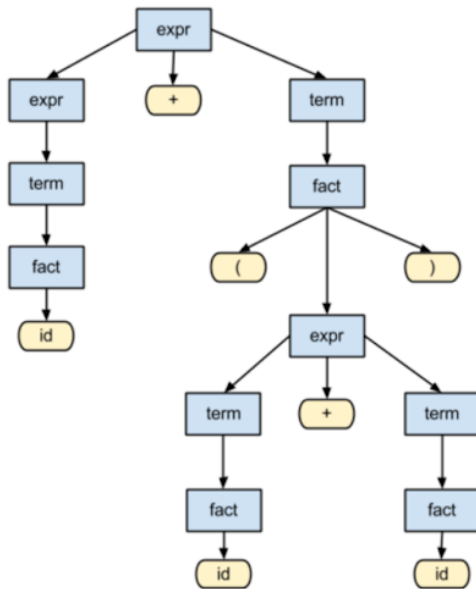
P ⇔ Reglas de producción

S ⇔ Axioma inicial

Analizador sintáctico.- Un analizador sintáctico o parser es una programa que normalmente es parte de un compilador y analiza una secuencia de caracteres según las reglas de una gramática formal. Un analizador léxico crea tokens de una secuencia de caracteres de entrada y son estos tokens los que son procesados por el analizador sintáctico para construir la estructura de datos, por ejemplo un árbol de análisis o árboles de sintaxis abstracta. La tarea del analizador es, en este caso, la descomposición y transformación de las entradas en un formato utilizable para su posterior procesamiento. Se analiza una cadena de instrucciones en un lenguaje de programación y luego se descompone en sus componentes individuales.

Funcionamiento.- Para analizar un texto, los analizadores suelen utilizar un analizador léxico separado (llamado lexer), que descompone los datos de entrada en fichas (token). Los Lexers son por lo general máquinas finitas, que siguen la gramática regular y por lo tanto aseguran un desglose adecuado. Los tokens obtenidos de esta manera sirven como caracteres de entrada para el analizador sintáctico.

El analizador actual maneja la gramática de los datos de entrada, realiza un análisis sintáctico de éstos y como regla general crea un **árbol de sintaxis (árbol de análisis)**. Esto se puede utilizar para el procesamiento posterior de los datos, por ejemplo, la generación de código por un compilador o ejecutado por un intérprete (traductor). Por lo tanto, el analizador es el software que comprueba, procesa y envía las instrucciones del código fuente.



Aplicación.- Un analizador sintáctico se utiliza a menudo para convertir texto en una nueva estructura, por ejemplo, un árbol sintáctico, que expresa la disposición jerárquica de los elementos. En las siguientes aplicaciones el uso de un analizador es usualmente esencial:

- La lectura de un lenguaje de programación es realizada por un analizador. Proporciona una estructura de datos al compilador, con la que se puede generar el código máquina o bytecode.
- El código HTML es al principio sólo una cadena de caracteres para un ordenador que debe ser analizada por el analizador contenido en el navegador web.
- Los motores de búsqueda como Google extraen (analizan) texto relevante para ellos de las páginas web descargadas con rastreadores. Se procesan y los datos analizados se pueden utilizar para la navegación.

Sintaxis y Gramática.- La sintaxis se define como la disposición de palabras en una oración para mostrar su relación. La sintaxis suministra información significativa que se necesita para entender un programa y proporciona información imprescindible para la traducción del programa fuente a un programa objeto.

La sintaxis tiene como principal función analizar el orden correcto de las palabras a fin de que las frases, oraciones, textos e ideas sean expresados de manera correcta para que pueda llegar el mensaje que se desea transmitir, esta suele ser más útil al usuario del lenguaje que al sistema de PLN.

Una gramática G es un modelo lingüístico-matemático que describe el orden sintáctico que deben cumplir las frases bien formadas de un lenguaje. Una gramática se define formalmente de la siguiente forma:

$G = (V_t, V_N, P, S)$ donde:

V_T : conjunto finito de símbolos terminales del lenguaje

VN :conjunto finito de símbolos no terminales

P : conjunto finito de reglas de producción

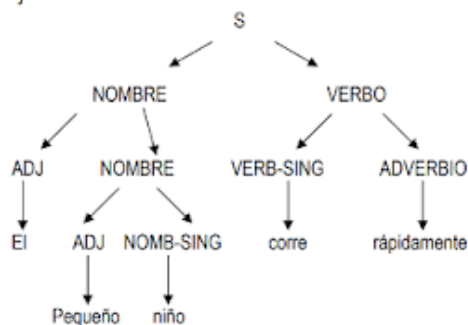
S : Símbolo distinguido o axioma inicial a partir del cual se reconocerán las secuencias de L aplicando sucesivamente las reglas de producción.

Consideremos la siguiente gramática G (V_N , V_T , S, P)
donde

$V_N = \{S, \text{NOMBRE}, \text{VERBO}, \text{ADJ}, \text{NOMB-SING}, \text{VERBO-SING}, \text{ADVERBIO}\}$

$V_T = \{\text{El}, \text{La}, \text{Los}, \text{Las}, \text{Pequeño}, \text{traviesa}, \text{niño}, \text{niña}, \text{estudia}, \text{corre}, \text{juega}, \text{salta}\}$

P = {S → NOMBRE VERBO
NOMBRE → ADJ NOMBRE
NOMBRE → ADJ NOMB-SING
VERBO → VERB-SING ADVERBIO
ADJ → El /La /Los /Las /Ellos /Ellas
ADJ → Pequeño /traviesa/ inquieto
NOMB-SING → niño / niña
VERB-SING → estudia / corre / juega /salta
ADVERBIO → rápidamente / despacio / mucho
}



Como podemos observar la frase formada a partir del lenguaje es: 'El Pequeño niño corre rápidamente' $\in L(G)$.

Durante el análisis sintáctico, se realizan derivaciones (de reglas gramaticales) a partir de un símbolo distinguido, para verificar si una frase pertenece al lenguaje definido por la gramática. A este proceso para determinar si es gramaticalmente correcta o no se le conoce como análisis sintáctico (parsing). Los árboles de análisis sintáctico muestran la sintaxis concreta de un lenguaje. Sin embargo, para aplicar esta gramática de forma mecánica y automatizada a una oración, es necesario contar con un lexicón que ofrezca información al analizador sintáctico (parser) sobre las categorías gramaticales que están asociadas a las palabras que aparecen en la oración que se desea analizar.

6. Aplicaciones de PLN

- Reconocimiento de patrones de lenguaje

Al procesar grandes cantidades de documentos, el reconocimiento de patrones permite filtrar datos importantes en cadenas de texto con un tiempo récord. Es el primer paso para la recuperación de información y para que la clasificación de textos sea posible.

- Recuperación de información

Gracias a que reconoce patrones de lenguaje, hace sencilla la tarea de encontrar un fragmento en particular dentro de una gran cantidad de texto. No inventa palabras o frases nuevas, sino que identifica la información valiosa.

- Traducciones automáticas de idiomas

Ya sea con voz o texto, esta función utiliza datos que se procesan por la lingüística computacional y está en un proceso constante de mejora y aprendizaje.

- Clasificación de información

Gracias a la aplicación de palabras clave, la información puede categorizarse para que su consulta sea más eficiente.

- Resumen de textos

Resumir un documento de gran extensión se apoya en ciertas palabras o frases clave. Seleccionando siempre la información más importante.

- Generación de lenguaje natural

Permite que una máquina responda las interacciones de un humano con frases nuevas.

- Síntesis de voz

Además de utilizarse para las instrucciones de voz que haces a un asistente virtual o tu computadora, traduce un texto a un audio que reconoce la separación de palabras, pausas que los signos gramaticales añaden a un discurso y, poco a poco, hasta la entonación adecuada.

7. Conclusiones

El procesamiento del lenguaje natural es un paso más en el camino de hacer que una computadora imite el comportamiento humano y nos ayuda a generar una comunicación con una persona de manera eficiente y entendible, esta comunicación está llegando tan lejos que actualmente ciertas herramientas y aplicaciones pueden leer y escribir en un lenguaje natural de forma tan realista que parece realizado directamente por una persona. Gracias al PLN un agente puede relacionar las palabras con objetos, acciones y circunstancias. de esta manera una máquina podría entender perfectamente una orden.