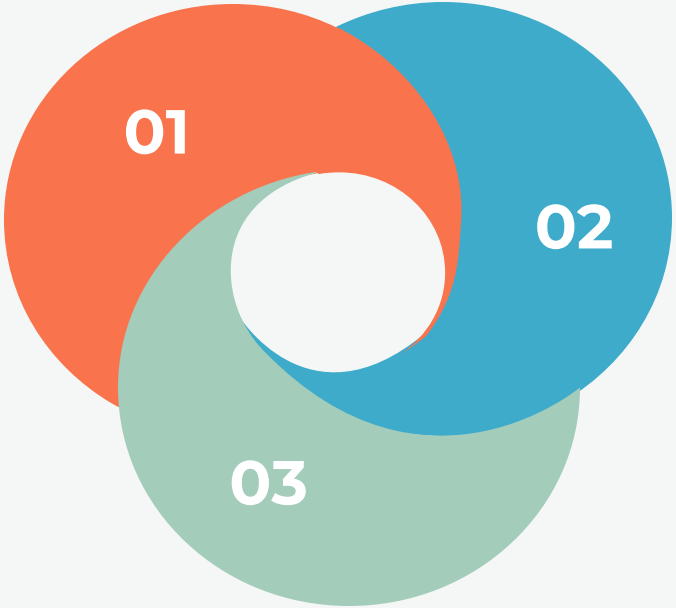
Introducción al Procesamiento del Lenguaje Natural

¿Qué es el Procesamiento del Lenguaje Natural?

El Procesamiento del Lenguaje Natural (PLN) es un subcampo de la Inteligencia Artificial que se centra en la interacción entre los ordenadores y el lenguaje humano.

Las técnicas de NLP se pueden aplicar a una amplia gama de tareas, como el análisis de sentimientos, la traducción automática, la recuperación de información y la generación de texto.



Implica el desarrollo de algoritmos y modelos que permiten a las computadoras comprender, interpretar y generar el lenguaje humano.

¿Por qué es difícil el procesamiento del lenguaje? (1/2)

Complejidad del lenguaje humano: El lenguaje humano es irregular, lleno de ambigüedades y matices. Las reglas gramaticales no siempre son rígidas, y las expresiones pueden variar según el contexto y la cultura.

Variabilidad en la forma de hablar: Las personas hablan con diferentes acentos, entonaciones y velocidades. Esto dificulta la creación de modelos precisos para reconocer y comprender el habla.

¿Por qué es difícil el procesamiento del lenguaje? (2/2)

Ambigüedad: Las palabras y frases pueden tener múltiples significados según el contexto. El PLN debe lidiar con esta ambigüedad para interpretar correctamente el significado deseado.

En resumen, el PLN combina la lingüística computacional con modelos estadísticos y de aprendizaje automático para que las máquinas puedan entender, generar y procesar texto y voz de manera efectiva.

Algunas aplicaciones del PLN

La siguiente lista no está completa, pero se han creado sistemas útiles para:

- Revisión ortográfica y gramatical
- Reconocimiento óptico de caracteres (OCR)
- Lectores de pantalla para usuarios ciegos y deficientes visuales
- Comunicación aumentativa y alternativa (es decir, sistemas para ayudar a las personas que tienen dificultades para comunicarse debido a una discapacidad)
- Traducción asistida por máquina (es decir, sistemas que ayudan a un traductor humano, por ejemplo, almacenando traducciones de frases)
- Herramientas para lexicógrafos
- Recuperación de información
- Clasificación de documentos (filtrado, enrutamiento)
- Agrupación de documentos
- Extracción de información
- Respuesta a preguntas
- Resumen
- Segmentación de texto
- Calificación de exámenes
- Generación de informes (posiblemente multilingüe)
- Traducción automática
- Interfaces de lenguaje natural para bases de datos
- Comprensión de correos electrónicos
- Sistemas de dálogo

Terminología de PLN

En el campo del Procesamiento del Lenguaje Natural (NLP), la **tokenización** es el proceso de dividir el texto en unidades más pequeñas, conocidas como tokens. Estos tokens suelen ser palabras, pero también pueden incluir signos de puntuación y otros elementos. La tokenización es un paso esencial porque prepara los datos para etapas posteriores de análisis y modelado. Por ejemplo, la frase "Los gatos duermen" se tokenizaría en "Los", "gatos" y "duermen". La **sintaxis**, por otro lado, se refiere a la estructura gramatical de las oraciones, y en NLP, se analiza para comprender cómo se organizan las palabras para transmitir significado. Otros términos comunes en NLP incluyen el **análisis morfológico**, que estudia la estructura de las palabras y cómo se forman; la **semántica**, que se ocupa del significado de las palabras y frases; y la **pragmática**, que examina cómo el contexto influye en la interpretación del lenguaje. Además, el **etiquetado de partes del discurso** es el proceso de asignar etiquetas a cada token para identificar su función gramatical, como sustantivo, verbo, adjetivo, etc. Estos conceptos son fundamentales para desarrollar aplicaciones de NLP que van desde los correctores ortográficos hasta los sistemas de traducción automática y los asistentes virtuales inteligentes.

Historia de PLN (1/3)

Los primeros esfuerzos en el campo del PLN se centraron en la creación de sistemas capaces de traducir textos de un idioma a otro. El experimento de Georgetown en 1954 fue uno de los primeros hitos, donde se utilizó un pequeño vocabulario y reglas simples para traducir frases del ruso al inglés. A pesar del optimismo inicial, los investigadores pronto se dieron cuenta de que la traducción automática era un problema mucho más complejo de lo que se había anticipado.

En los años siguientes, el enfoque reglamentado continuó dominando el campo. Los lingüistas y científicos de la computación intentaban codificar el conocimiento del lenguaje en forma de gramáticas y diccionarios. La teoría de la Gramática Generativa de Noam Chomsky, que introdujo la idea de una gramática universal subyacente a todos los idiomas humanos, tuvo una influencia significativa en este período. Los sistemas de PLN de la época intentaban modelar el lenguaje utilizando estas reglas universales, pero se encontraban con la dificultad de capturar todas las sutilezas y excepciones del lenguaje natural.

Durante los años 60 y 70, surgieron los primeros sistemas conversacionales como ELIZA y SHRDLU. ELIZA, creado por Joseph Weizenbaum, fue diseñado para simular una conversación y podía adoptar diferentes roles, como el de un psicoterapeuta. SHRDLU, desarrollado por Terry Winograd, permitía a los usuarios interactuar con un mundo virtual de bloques a través de comandos en lenguaje natural. Estos sistemas eran impresionantes para su tiempo, pero estaban limitados por su incapacidad para comprender realmente el lenguaje más allá de las interacciones predefinidas y los contextos muy específicos.

Historia de PLN (2/3)

El PLN también se benefició de los avances en otras áreas de la inteligencia artificial, como la representación del conocimiento y la comprensión de la semántica. Los investigadores exploraron cómo las computadoras podrían no solo procesar la estructura del lenguaje, sino también entender su significado. Sin embargo, estos sistemas semánticos eran difíciles de escalar y requerían una gran cantidad de trabajo manual para codificar el conocimiento del mundo.

A finales de los años 70 y principios de los 80, el campo del PLN comenzó a experimentar con enfoques estadísticos. Estos métodos, aunque todavía en su infancia, prometían una forma de procesar el lenguaje que no dependiera de reglas rígidas y que pudiera aprender de grandes cantidades de datos. Sin embargo, no fue hasta la llegada del aprendizaje automático y el aumento del poder de cómputo que estos enfoques estadísticos pudieron ser plenamente aprovechados.

Historia de PLN (3/3)

El procesamiento de lenguaje natural (PLN) ha experimentado una transformación significativa en la última década, marcada por avances tecnológicos que han redefinido las posibilidades de interacción entre humanos y máquinas. Modelos como BERT (Bidirectional Encoder Representations from Transformers) y GPT (Generative Pre-trained Transformer) han sido fundamentales en esta evolución. BERT, desarrollado por Google, introdujo un enfoque revolucionario en el preentrenamiento de representaciones lingüísticas, permitiendo que el modelo comprenda el contexto de una palabra en relación con todas las otras palabras en una oración, no solo las precedentes o subsecuentes. Esto ha mejorado significativamente la comprensión del lenguaje y ha tenido un impacto considerable en tareas como la clasificación de texto y la respuesta a preguntas.

Por otro lado, la serie de modelos GPT de OpenAI, comenzando con GPT-1 en 2018, ha llevado la generación de texto a nuevos niveles de fluidez y coherencia. GPT-2 y GPT-3, con su número creciente de parámetros, han mostrado una habilidad impresionante para generar texto que puede ser indistinguible del escrito por humanos en ciertas tareas. GPT-4 ha marcado otro hito, con mejoras en la comprensión de contextos complejos y la capacidad de interactuar en modalidades multimodales, aceptando imágenes y generando código a partir de ellas. Este avance ha abierto nuevas fronteras en la interacción humano-computadora, permitiendo que las máquinas comprendan y respondan a entradas más allá del texto tradicional.

¡Gracias por su atención! 🙂