

Capítulo 6

Una introducción al procesamiento de
lenguaje natural

El procesamiento del lenguaje natural

- El **procesamiento del lenguaje natural** (**PLN**) es una subdisciplina de la inteligencia artificial (**IA**) que se ocupa de la **comprensión**, **interpretación** y **generación** del lenguaje humano por parte de las máquinas.
 - Su **objetivo** es permitir que las **computadoras** comprendan el **texto** y el **habla** de **manera similar** a como lo haría un **ser humano**

El procesamiento del lenguaje natural

- **Componentes Clave:**

1. Análisis Sintáctico: Estudia la estructura gramatical de las oraciones, identificando partes del discurso (sustantivos, verbos, etc.) y su relación.

- **Ejemplo:** Consideremos la oración: "El banco me aprobó el préstamo". La estructura gramatical identifica "El banco" como el sujeto, "me aprobó" como el verbo y "el préstamo" como el objeto directo.

2. Análisis Semántico: Se enfoca en el significado de las palabras y frases en contexto, ayudando a las máquinas a entender el contenido.

- **Ejemplo:** en el ejemplo anterior "banco" puede referirse a una institución financiera o a un banco (asiento). El análisis semántico utiliza el contexto (la relación con "aprobar" y "préstamo") para discernir que se refiere a una institución financiera.

El procesamiento del lenguaje natural

- **Componentes clave (cont):**

1. **Reconocimiento de Entidades Nombradas (NER):** Identifica y clasifica entidades dentro del texto, como nombres de personas, lugares o fechas.
2. **Generación de Lenguaje Natural (NLG):** Permite a las máquinas generar texto coherente y relevante a partir de datos estructurados.
3. **Análisis de Sentimientos:** Evalúa el tono emocional del texto, determinando si es positivo, negativo o neutral.

El procesamiento del lenguaje natural

- **Aplicaciones Comunes:**

- **Asistentes Virtuales:** Como Siri o Alexa, que utilizan NLP para comprender y responder a comandos de voz.
- **Traducción Automática:** Herramientas como Google Translate que convierten texto de un idioma a otro.
- **Chatbots:** Sistemas que interactúan con los usuarios en tiempo real, brindando soporte al cliente o información.
- **Análisis de Opiniones:** Utilizado por empresas para evaluar la percepción del cliente sobre productos o servicios a través de comentarios en redes sociales.

El procesamiento del lenguaje natural

- La comprensión del lenguaje natural (NLU) y la generación de lenguaje natural (NLG) son subconjuntos importantes del PLN.
- La **Comprensión del Lenguaje Natural (NLU)** es un **subcampo** de la **inteligencia artificial** centrado en capacitar a las máquinas para **comprender** e **interpretar** el **lenguaje humano** de forma significativa.
 - A diferencia del simple reconocimiento de palabras, el NLU trata de entender el **contexto**, la **intención** y el **sentimiento** que hay detrás del lenguaje utilizado.

El procesamiento del lenguaje natural

- **Ejemplos de aplicaciones cotidianas de NLU:**

- Un **asistente virtual** usa NLU para comprender comandos de voz y preguntas.
- Los **chatbots** usan NLU para interactuar con los usuarios.
- Otra aplicación de NLU es en el **análisis de sentimiento** donde se analizan comentarios en redes sociales, reseñas de productos y encuestas; se puede determinar si el texto analizado es positivo, negativo o neutral.
- Otro uso de NLU es la **clasificación de correos electrónicos** (por ejemplo: si es spam o no; o clasificar correos como urgentes, promociones o reuniones.)
- Además, se puede usar NLU para **comprender el texto** en un idioma y **traducirlo** a otro idioma.
- **Programas de procesamiento de texto** utilizan NLU para identificar errores gramaticales y ofrecer sugerencias.

El procesamiento del lenguaje natural

- La generación de lenguaje natural (NLG) es un subcampo de la inteligencia artificial que se centra en la creación automática de narraciones escritas o habladas a partir de datos estructurados.
 - Permite a las máquinas generar textos coherentes, contextualmente relevantes y parecidos al lenguaje humano.
- El NLG puede aplicarse en diversos ámbitos, como la generación de informes, las respuestas de atención al cliente y la creación de contenidos.
- Ahora vemos las etapas de NLG.
 - Fueron extraídos de sitios web recomendados por perplexity (entre ellos wikipedia).

Etapas de la comprensión del lenguaje natural

1. **Preprocesamiento:** Consiste en limpiar y preparar los datos de texto para el análisis.
 - Contempla tareas como la **tokenización** (descomposición del texto en unidades más pequeñas – p.ej. Palabras y frases), **etiquetado de parte del discurso** (identificación de funciones gramaticales de las palabras), **análisis sintáctico** (de la estructura gramatical de las frases).
2. **Análisis semántico:** esta etapa se centra en comprender el **significado** de las palabras en su **contexto**.
 - Contempla tareas como el **reconocimiento de entidades** (identificación y clasificación de entidades clave en el texto – p.ej. Nombres, lugares, fechas, etc.), **extracción de relaciones** entre las diferentes **entidades** del texto.

Etapas de la comprensión del lenguaje natural

3. **Comprendiendo el contexto:** Se analiza el contexto más amplio de una conversación o texto.

- Contempla tareas como :
 - evaluar el **tema general** y los matices que afectan a la interpretación,
 - reconocer **significados específicos del contexto** que pueden cambiar en función de las **interacciones anteriores** o del **texto circundante**.

4. **Reconocimiento de intenciones:** Determina la intención del **usuario** tras un texto o discurso determinado.

- Esto es crucial para aplicaciones como **chatbots** y **asistentes virtuales**, ya que ayuda a proporcionar respuestas adecuadas.
- **Ejemplos de intención:** obtener información sobre un tema específico, resolver una duda o problema específico, conversar por curiosidad o entretenimiento, realizar una acción concreta (traducir, resumir, etc.)

Etapas de la generación del lenguaje natural

1. **Determinación de contenido:** se decide qué información **debe incluirse** en el texto generado.
 - La selección de los **datos relevantes** garantiza que el **resultado tenga sentido** y se ajuste a las necesidades del usuario o a tareas específicas.
 - **Ejemplo:** Al generar un **informe meteorológico**, el sistema podría determinar incluir la **temperatura**, la **humedad** y las **previsiones**.
2. **Estructuración de documentos:** Se organiza la información de manera lógica.
 - Se estructura el flujo de información en **párrafos** y **secciones**.
 - Se deciden los **títulos** o **viñetas** para mayor claridad.
 - Un documento bien estructurado mejora la legibilidad y la comprensión.

Etapas de la generación del lenguaje natural

3. **Planificación de frases:** el sistema planifica cómo expresar en frases el contenido seleccionado.
 - Involucra elegir las **estructuras gramaticales** adecuadas.
 - También contempla decidir el uso de **pronombres** y **conjunciones** para dar coherencia.
 - Esta etapa garantiza que las oraciones generadas sean gramaticalmente correctas y fluidas.
4. **Generación de expresiones de referencia:** se trata de crear expresiones que hagan **referencia** a **entidades** o **conceptos** mencionados en el texto.
 - La correcta referencia ayuda a mantener la claridad y a evitar la ambigüedad.

Etapas de la generación del lenguaje natural

5. **Realización:** Las frases planificadas se transforman en texto real.

- Se aplican reglas sintácticas, morfológicas y ortográficas.
- Se garantiza la fluidez y la coherencia del texto generado.
- Esta etapa produce el texto final que leerán los usuarios, que debe ser pulido y profesional.

6. **Post procesamiento (opcional):** se produce un refinamiento adicional luego de la generación inicial del texto.

- Se comprueban errores gramaticales o frases incómodas.
- Se mejora el estilo o el tono en función de las preferencias del usuario o del contexto.
- El post procesamiento puede mejorar aún más la calidad del resultado.

Etapas del preprocesamiento

1. **Limpieza de datos:** se eliminan elementos innecesarios del texto que pueden interferir con el análisis.
 - Se eliminan caracteres especiales (p.ej. puntuaciones, símbolos y caracteres no alfabéticos).
 - **Conversión a minúsculas:** esto permite uniformizar el texto para evitar duplicados (p.ej. “hola” y “Hola”)
 - Eliminación de etiquetas HTML si el texto viene de la web.
2. **Tokenización:** Se divide el texto en unidades más pequeñas llamadas tokens (p.ej. Palabras o frases)
3. **Eliminación de stop words:** elimina palabras que no aportan significado relevante al análisis, como “y”, “el”, “de”.
 - Así se reduce la cantidad de datos y mejora la eficiencia del procesamiento al enfocarse en palabras significativas.

Etapas del preprocesamiento

4. Stemming y lematización:

- **Stemming** es un proceso que corta las palabras a su raíz, eliminando sufijos y prefijos. P.ej: “cocinando” se convierte en “cocin”.
- **Lematización**: similar a stemming, pero devuelve la forma base o raíz de una palabra considerando su contexto gramatical. P.ej: “mejores” se convierte en “bueno”.

5. Análisis morfológico:

se centra en la estructura interna de las palabras, identificando sus componentes (raíces, sufijos).

- Ayuda a entender cómo se forman las palabras y sus significados.

Etapas del preprocesamiento

6. **Part of speech tagging:** Asigna etiquetas gramaticales a cada token (sustantivo, verbo, adjetivo, etc.)
 - Facilita la comprensión de la función de cada palabra dentro de una oración.
7. **Conversión a formato estructurado:** se transforma el texto preprocesado a un formato que pueda ser fácilmente analizado por algoritmos.
 - Se pueden extraer entidades nombradas.
 - Se puede organizar la información en formato estructurado (p.ej. Tabla de BD, o formatos para intercambio de datos como JSON o XML.)
 - Se puede traducir las frases y palabras a representaciones numéricas usando técnicas como Word2Vec, GloVe o FastText que convierten palabras en vectores densos que capturan relaciones semánticas entre ellas.

Ejemplo de preprocessamiento

- Consideramos la frase:
 - “Prosoft announced a new producto in Seattle on Monday”
- **Tokenization:**
 - [“Prosoft”, “announced”, “a”, “new”, “product”, “in”, “Seattle”, “on”, “Monday”]
- **Eliminación de stop words:**
 - [“Prosoft”, “announced”, “new”, “product”, “Seattle”, “Monday”]
- **Lemmatización:**
 - [“Prosoft”, “announce”, “new”, “product”, “Seattle”, “Monday”]

Ejemplo de preprocessamiento

- Part of speech tagging:
 - [("Prosoft", "NNP"), ("announce", "VBP"), ("new", "JJ"), ("product", "NN"), ("Seattle", "NNP"), ("Monday", "NNP")]
- Donde:
 - NNP: nombre propio.
 - VBP: Verbo
 - JJ: Adjetivo
 - NN: sustantivo singular

Etapas del Análisis Semántico

1. **Desambiguación de Palabras:** Este proceso implica determinar el significado correcto de una palabra que puede tener múltiples significados (polisemia) dependiendo del contexto.
 - **Ejemplo:** La palabra "banco" puede referirse a una entidad financiera o a un banco para sentarse. El análisis semántico ayuda a identificar cuál es el significado adecuado según el contexto.
2. **Reconocimiento de Entidades Nombradas (NER):** Consiste en identificar y clasificar entidades clave dentro del texto, como nombres de personas, lugares, organizaciones y fechas.
 - **Importancia:** Permite entender los detalles específicos y relevantes del contenido, facilitando la extracción de información significativa.
3. **Análisis de Relaciones:** Esta etapa se enfoca en identificar las relaciones entre diferentes entidades dentro del texto.
 - **Ejemplo:** En la frase "El presidente de Francia visitó Alemania", se establece una relación entre "presidente" y "Francia", así como entre "Francia" y "Alemania".

Etapas del Análisis Semántico

4. **Análisis de Sentimientos:** Se utiliza para determinar el tono emocional del texto, identificando si es positivo, negativo o neutral.
 - **Importancia:** Es especialmente útil en aplicaciones como el análisis de opiniones y la monitorización de redes sociales.
5. **Comprensión Contextual:** Implica analizar el contexto general en el que se presenta el texto, incluyendo el tema y las interacciones previas.
 - **Ejemplo:** Si un usuario ha mencionado anteriormente un proyecto específico, el sistema puede adaptar sus respuestas futuras basándose en ese contexto.
6. **Construcción de Representaciones Semántica:** Se refiere a la creación de modelos que representan el significado del texto en un formato estructurado que las máquinas pueden procesar.
 - **Importancia:** Estas representaciones permiten a los sistemas realizar inferencias y responder preguntas basadas en el contenido analizado.

Ejemplo de reconocimiento de entidades nombradas

- Partimos del resultado del ejemplo anterior:
 - [("Prosoft", "NNP"), ("announce", "VBP"), ("new", "JJ"), ("product", "NN"), ("Seattle", "NNP"), ("Monday", "NNP")]
- Luego de aplicar reconocimiento de entidades nombradas obtenemos:
 - [("Prosoft", "ORG"), ("announce", "O"), ("new", "O"), ("product", "O"), ("Seattle", "GPE"), ("Monday", "DATE")]
- Donde:
 - ORG: Organización.
 - O: otro (no es una entidad nombrada)
 - GPE: Entidad geopolítica (ciudad, país, etc.)
 - DATE: fecha

Ejemplo de análisis de relaciones

- **Análisis de relaciones:**

- **Prosoft** es el **sujeto** realizando la acción **announced**.
- **Announced** es el **verbo** actuando sobre el **objectIdo product**.
- **Product** es descrito como **nuevo** por el adjetivo.
- **Seattle** prove la **localización** donde el anuncio toma lugar.
- **Monday** especifica cuando se hace el anuncio.

Ejemplos de análisis del sentimiento

- **Ejemplo 1:** ¡Me acaban de ascender en el trabajo y me siento en la cima del mundo! Análisis del sentimiento:
 - Palabras positivas: ascendido, cima
 - Puntuación del sentimiento: Alto sentimiento positivo debido a las palabras que indican éxito y felicidad.
- **Ejemplo 2:** Estoy muy disgustada porque mis vacaciones se cancelaron en el último minuto. Análisis del sentimiento:
 - Palabras negativas: disgustada, cancelado
 - Puntuación del sentimiento: Alto sentimiento negativo debido a palabras que indican decepción y frustración.

El contexto

- En el marco del PLN, el término **contexto** se refiere a la información adicional que rodea una palabra, frase o conversación que ayuda a interpretar su significado y a entender la intención detrás de las interacciones.
- El contexto es crucial para desambiguar significados, mantener la coherencia en el diálogo y proporcionar respuestas relevantes.
- Comprender los **diferentes tipos de contexto** permite a los sistemas NLU gestionar diálogos complejos y proporcionar respuestas precisas y relevantes.

Tipos de contexto en PLN

1. **Contexto Local**: Se refiere al entorno inmediato de una palabra o frase dentro de una oración.
 - Incluye las palabras que preceden y siguen a la unidad de texto en cuestión.
 - **Ejemplo**: En la frase "El banco está cerrado", la palabra "banco" puede referirse a una institución financiera o a un banco de un río, dependiendo de las palabras que lo rodean.
2. **Contexto Global**: Abarca información que se extiende más allá de una sola oración o párrafo.
 - Incluye el historial de interacciones anteriores en un diálogo o el tema general de una conversación.
 - **Ejemplo**: Si un usuario pregunta "¿Dónde está mi pedido?" después de haber mencionado previamente un artículo específico, el sistema debe recordar esa información para proporcionar una respuesta precisa.

Tipos de contexto en PLN

3. **Contexto Conversacional:** Se refiere a la dinámica del diálogo entre los participantes, incluyendo turnos de habla y respuestas anteriores.
 - Es esencial para mantener la fluidez y la relevancia en las interacciones.
 - **Ejemplo:** En un diálogo donde un usuario dice "Me gustaría reservar una mesa", y luego sigue con "¿A qué hora?", el sistema debe entender que "hora" se refiere al momento para la reserva mencionada anteriormente.
4. **Contexto Situacional:** Incluye factores externos que pueden influir en la interpretación del lenguaje, como el entorno físico, el estado emocional del hablante o el contexto cultural.
 - **Ejemplo:** Si un usuario dice "¿Puedes pasarme eso?" mientras señala un objeto, el sistema debe entender que "eso" se refiere al objeto señalado, no a cualquier otro.

Tipos de contexto en PLN

5. **Contexto Histórico:** Se refiere al conocimiento acumulado a lo largo del tiempo sobre un tema o sobre las preferencias del usuario.
 - Esto es especialmente relevante en sistemas que interactúan repetidamente con los mismos usuarios.
 - **Ejemplo:** Un asistente virtual que recuerda preferencias pasadas, como "Siempre pido pizza vegetariana", puede personalizar mejor sus recomendaciones futuras.
6. **Contexto Cultural:** Incluye factores sociales y culturales que influyen en cómo se interpreta el lenguaje.
 - Este contexto puede afectar tanto la forma como el contenido de la comunicación.
 - **Ejemplo:** Frases coloquiales o expresiones idiomáticas pueden tener significados diferentes dependiendo del trasfondo cultural del hablante y del oyente.

Comprensión del contexto

- La **comprensión del contexto** es esencial para interpretar correctamente el significado de las entradas en lenguaje natural.
 - Esta capacidad permite a los sistemas NLU entender no solo las palabras individuales, sino también cómo se relacionan entre sí en un contexto más amplio.
 - Se centra en entender cómo el significado se ve afectado por el contexto en el que se presenta.
 - El contexto es fundamental para la interpretación precisa del lenguaje humano.
 - Las palabras y frases pueden tener significados diferentes según la situación, el tono y la información previa.
 - **Por ejemplo**, la frase "¿Puedes pasarme eso?" puede referirse a un objeto específico solo si se conoce el contexto de la conversación.

Comprensión del contexto

- La comprensión del **contexto conversacional** implica seguir el hilo de una conversación a lo largo del tiempo. Esto incluye recordar interacciones pasadas y utilizar esa información para influir en las respuestas futuras.
- Algunos sistemas NLU implementan **memoria contextual**, donde retienen información de **interacciones previas**.
 - Esto es especialmente útil en aplicaciones como chatbots y asistentes virtuales, donde las conversaciones pueden ser continuas.
 - **Ejemplo:** Si un usuario menciona un proyecto específico en una conversación anterior, el sistema puede referirse a ese proyecto en interacciones futuras, mejorando así la relevancia de las respuestas.

Comprensión del contexto

- El **contexto situacional** se refiere a factores externos que pueden influir en la interpretación del lenguaje. Esto incluye aspectos como la ubicación geográfica, la hora del día y eventos actuales.
 - **Ejemplo Práctico:** Si un usuario pregunta "¿Dónde puedo comprar café?" durante la mañana en una aplicación de asistente virtual, el sistema podría inferir que se refiere a cafeterías cercanas y no a tiendas de comestibles que venden café en grano.
- La comprensión del contexto también abarca el **análisis de sentimientos**, que permite al sistema captar las emociones detrás de las palabras. Esto es crucial para adaptar las respuestas según el tono emocional del usuario.
 - **Importancia:** Un análisis efectivo del sentimiento puede ayudar a determinar si un usuario está satisfecho o frustrado, permitiendo al sistema ajustar su enfoque y ofrecer soluciones más adecuadas.

Comprensión del contexto

- La capacidad de **adaptación dinámica** es esencial para mejorar la interacción entre humanos y máquinas.
 - Esto quiere decir que los sistemas NLU deben ser capaces de ajustar sus respuestas en tiempo real basándose en el contexto proporcionado por los usuarios.
 - **Ejemplo Práctico:** Si un usuario cambia repentinamente de tema durante una conversación, un sistema eficaz debería reconocer este cambio y adaptar sus respuestas para alinearse con el nuevo tema sin perder coherencia.

Etapas de Comprensión del contexto

- 1. Análisis Sintáctico:** Implica descomponer una oración en sus componentes gramaticales, como sujeto, verbo y objeto, para entender su estructura.
 - Permite identificar cómo se combinan las palabras y cómo su disposición afecta el significado.
 - Esto es fundamental para establecer relaciones entre las palabras dentro de la oración.
- 2. Desambiguación de Palabras:** Se refiere a la identificación del significado correcto de palabras que pueden tener múltiples significados dependiendo del contexto.
 - **Ejemplo:** La palabra "banco" puede referirse a una institución financiera o a un banco para sentarse. El sistema debe discernir cuál es el significado correcto según el contexto.

Etapas de Comprensión del contexto

3. **Análisis de Relaciones:** Esta etapa se centra en entender cómo las diferentes entidades y conceptos dentro del texto están relacionados entre sí.
 - **Ejemplo:** En "El presidente de Francia visitó Alemania", se establece una relación entre "presidente" y "Francia", así como entre "Francia" y "Alemania".
4. **Análisis Pragmático:** considera el contexto situacional, la intención del hablante y las implicaciones sociales.
 - Permite interpretar correctamente el lenguaje en situaciones específicas, teniendo en cuenta factores como el tono, las convenciones conversacionales y las expectativas culturales.

Etapas de Comprensión del contexto

5. **Resolución de Correferencia:** tarea de identificar a qué entidades se refieren los pronombres o expresiones en un texto.
 - Es crucial para mantener la coherencia en un diálogo, ya que permite al sistema recordar a quién o qué se refiere cuando se utilizan términos como "él", "ella" o "eso".
6. **Comprensión Contextual:** implica analizar el contexto más amplio en el que se presenta el texto, incluyendo interacciones previas y el tema general de la conversación.
 - Permite adaptar las respuestas según la información ya conocida sobre el usuario o el tema discutido.
 - Incluye el mantenimiento del contexto conversacional.
 - Algunos sistemas NLU mantienen un registro (llamado memoria contextual) de interacciones anteriores para mejorar la comprensión en conversaciones continuas.
 - **Ejemplo:** Si un usuario menciona un proyecto específico en una conversación anterior, el sistema puede hacer referencia a ese proyecto en interacciones futuras.

Etapas de Comprensión del contexto

7. **Análisis de Sentimientos:** Determina el tono emocional del texto, identificando si es positivo, negativo o neutral.
 - Es útil para comprender la actitud del usuario hacia un tema específico y ajustar las respuestas en consecuencia.
8. **Adaptación al Contexto Cultural:** Considera las variaciones culturales y lingüísticas que pueden influir en la interpretación del lenguaje.
 - Ayuda a los sistemas a adaptarse a diferentes expresiones idiomáticas y convenciones lingüísticas que pueden variar entre diferentes grupos culturales.