Procesamiento de Lenguaje Natural

Fabián Villena

Introducción

El NLP (Natural Language Processing) es una disciplina de la intersección entre la computación y la lingüística cuyo objeto de estudio es la habilidad de los computadores para manipular texto en lenguaje natural.

```
paciente de 71 años, con
antecedentes de hta en tto,
diabetes insulinodependiente,
dislipidemia, hipotiroidismo
en tto, enfermedad renal
cronica etapa iii,tabaquismo
cronico importante, en
febrero de este año lo
suspendio. Refiere que tiene
principios de Alzheimer y
parkinson?????? NO SALE
REGISTRO DE DIAGNOSTICOS.
```

Datos

Los datos se pueden dividirse en dos tipos:

1. Datos estructurados donde conocemos el modelo de datos subyacente
2. Datos no estructurados donde no existe un modelo de datos y presentan alta ambigüedad e irregularidades.

```
Paciente:
    id: <número>
    Nombre: <texto>
    F. de Nac.: <fecha>
    Salario: <número>
    Caries: <número>


id  Nombre ...    Caries
1   Juan   ...    10
2   Pedro  ...    8
...
```

Desafíos en el NLP

El texto es una medio de persistencia de datos muy ambiguo y poco sistematizado, por lo que la extracción de información desde estas fuentes es un desafío.

Existen significativas diferencias entre datos de texto generados en distintos dominios, por lo que un sistema diseñado para analizar texto de un dominio, puede que no funcione en un dominio distinto.

Abreviaciones

La abreviación es la forma de escribir un concepto sin usar el deletreo completo. Las abreviaciones son una forma eficiente de escribir pero ralentiza la lectura debido a que el lector debe interpretar la abreviación.

En algunos casos las abreviaciones pueden ser ambiguas, como en el caso de `OI` -> *ojo izquierdo*, *oído izquierdo*.

- HTA: Hipertensión arterial
- Pza.: Pieza
- ERC: Enfermedad renal crónica
- DMII: Diabetes mellitus tipo dos
- Tu.: Tumor
- Neo.: Neoplasia

Afirmaciones

Las afirmaciones son preposiciones que pueden tener algún nivel de polaridad positiva o negativa. Pueden ir desde negaciones completas, pasando por especulaciones hasta llegar a afirmaciones completas.

- Paciente de 6 años **libre de** caries.
- La **madre refiere que la paciente no** ha tenido fiebre.
- Se recibe paciente derivado desde el cesfam con **sospecha de** cacu.
- Paciente de 4 años **con** caries en diente 5.1.
- El paciente presenta una **franca** desviación de la línea media incisiva

Malformaciones

Debido a que en ciertas situaciones los profesionales de la salud cuentan con un tiempo acotado para volcar la información a la ficha clínica, el texto contenido en el registro clínico tiende a presentar malformaciones como faltas de ortografía, y presencia de inserciones, deleciones o sustituciones de caracteres.

PCTE CON CUADROS DE PERICORONITIS RECURRENTE EN ~~la~~ ZONA ~~de~~ ~~la~~ PZA 3.8 SEMIERUPCIONADA, SE RUEGA EVALUACION PARA EVETNUAL CIRUGIA DE EXODONCIA ~~de~~ PZA 3.8 Y POSIBLEMENTE ~~de~~ PZA 4.8

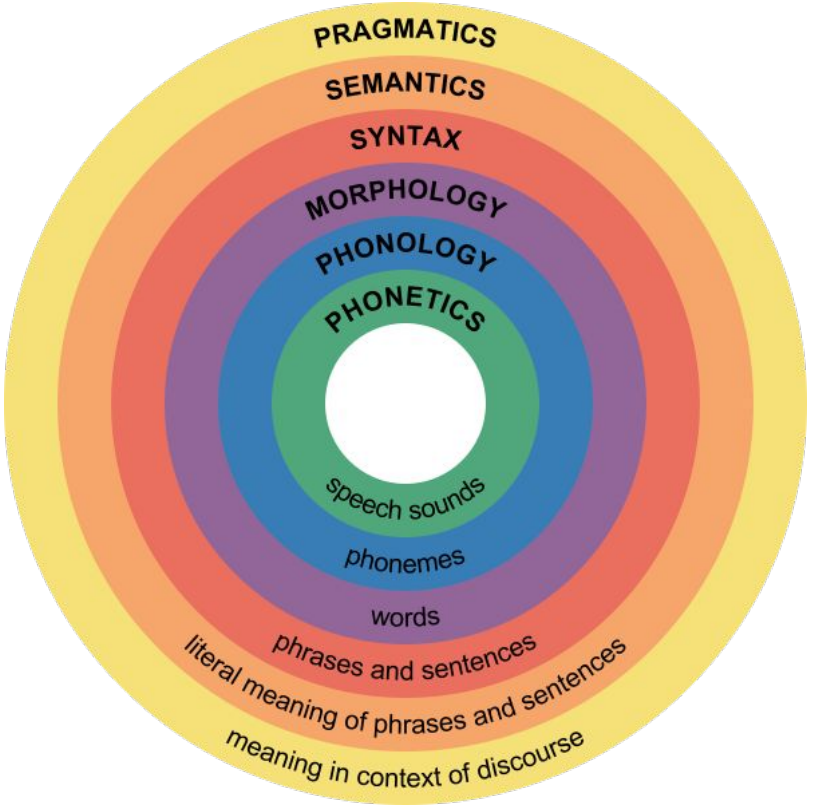
Generación de recursos lingüísticos

Se necesitan muestras de ejemplo de narrativas de lenguaje natural para poder adaptar un sistema de NLP, por lo que se necesita obtener y enriquecer grandes volúmenes de texto para poder ajustar un modelo de NLP.

La generación de recursos lingüísticos para NLP es un proceso altamente costoso dada la complejidad del enriquecimiento de grandes volúmenes de texto por parte de humanos expertos.

Niveles de análisis en NLP

- Fonética: Estructura de los sonidos.
- Morfología: Significados en partes de las palabras.
- Sintaxis: Estructura de una oración.
- Semántica: Significado de oraciones.



Fonología

La fonética es el estudio de lenguaje a nivel de los sonidos y la fonología es el estudio de la combinación de sonidos en unidades organizadas de discurso.

Este nivel analiza la interpretación de los sonidos de los discursos a través de las palabras. En sistemas de NLP que reciben entradas de sonido, las ondas de sonido son analizadas y representadas en una señal digital para que esta sea interpretada por el sistema de NLP.

Morfología

La morfología es el estudio de las estructuras y funciones internas de las palabras y cómo las palabras están formadas por unidades más pequeñas llamadas morfemas.

Para muchas tareas de NLP es útil remover la morfología inflexional en donde no se cambia el significado de la palabra pero se disminuye el tamaño del vocabulario.

Sintáxis

El análisis sintáctico es el proceso de análisis del lenguaje natural con las reglas de una gramática formal. Las reglas gramaticales son aplicadas a categorías o grupos de palabras, no a palabras individuales.

Para realizar el análisis sintáctico se necesita un parser, el cual es un componente que toma texto de entrada y provee una representación estructural de los datos después de una validación de una sintaxis correcta.

Semántica

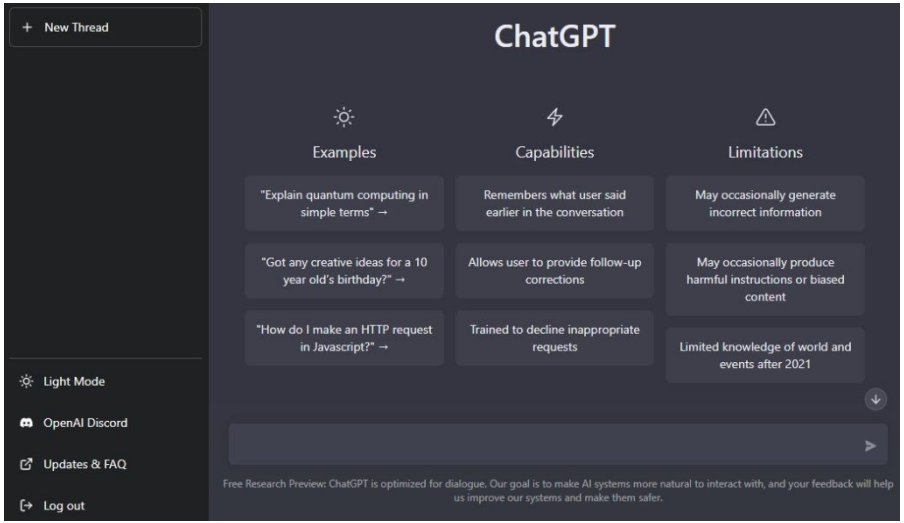
La semántica es el estudio de la referencia, significado o verdad. El análisis semántico es el proceso del entendimiento del lenguaje natural al extraer información relevante desde datos no estructurados.

Este nivel de análisis es uno de los más importantes en NLP y es donde el área de aprendizaje de representaciones centra su atención para lograr representar el significado de piezas de texto a través de vectores numéricos.

Técnicas de análisis de texto

Existen múltiples técnicas para analizar texto, desde métodos basados en reglas hasta modelos ajustados con algoritmos de aprendizaje automático.

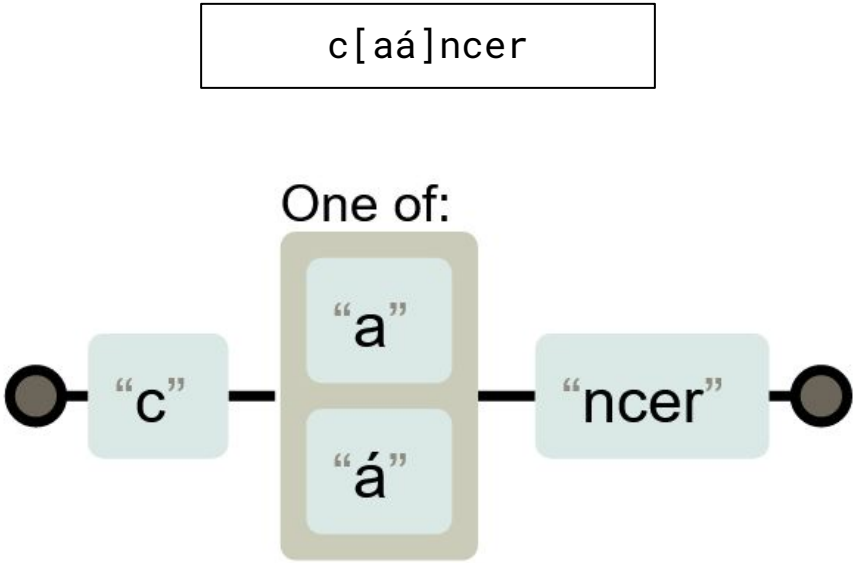
Todos los métodos modernos y del estado del arte están basados en aprendizaje automático. Al complejizar los métodos de análisis, perdemos la explicabilidad del sistema, que en medicina puede ser especialmente importante.



Métodos basados en reglas

Estos métodos utilizan el paradigma clásico de programación donde un desarrollador escribe reglas para imitar el comportamiento requerido del programa.

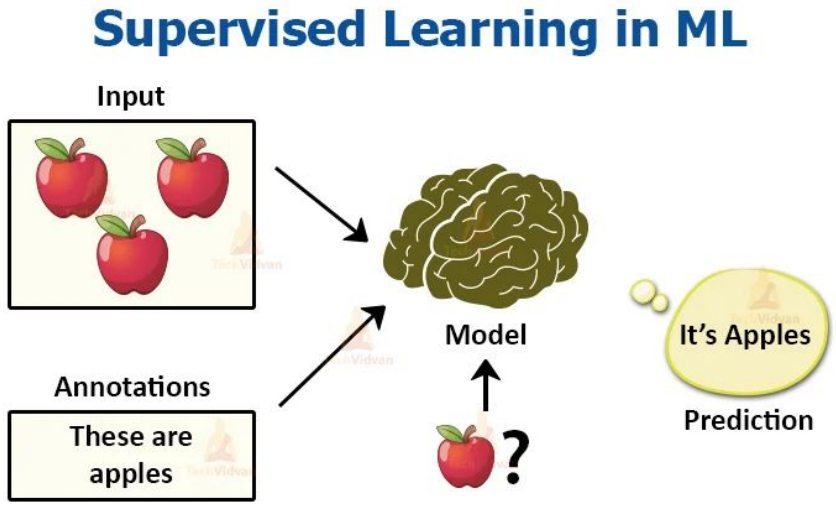
Lo más utilizado son las expresiones regulares para encontrar patrones dentro de cadenas de texto.



Métodos de aprendizaje automático

Con los métodos de aprendizaje automático, no se necesita un desarrollador para escribir a mano las reglas que queremos seguir para detectar lo que necesitamos.

A través de un algoritmo aprendemos esas reglas para generar una función que modela el proceso.



Ingeniería de características

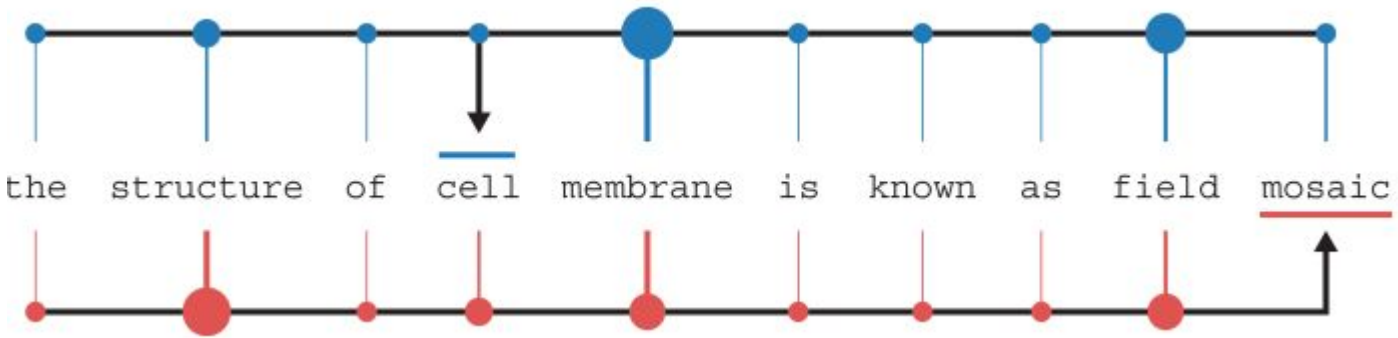
La ingeniería de características consiste en la extracción y transformación de variables desde datos crudos. La ingeniería de características es difícil porque necesita conocimiento de análisis de datos, conocimiento del dominio y además intuición.

Se necesita una alta experticia en el dominio para poder extraer y transformar las variables que mejor representan los datos no estructurados.

Deep Learning

A través de la utilización de complejas redes neuronales artificiales somos capaces de modelar el lenguaje de una manera muy precisa.

Con el advenimiento de las arquitecturas basadas en autoatención llamadas Transformers han aparecido sorprendentes sistemas de NLP.

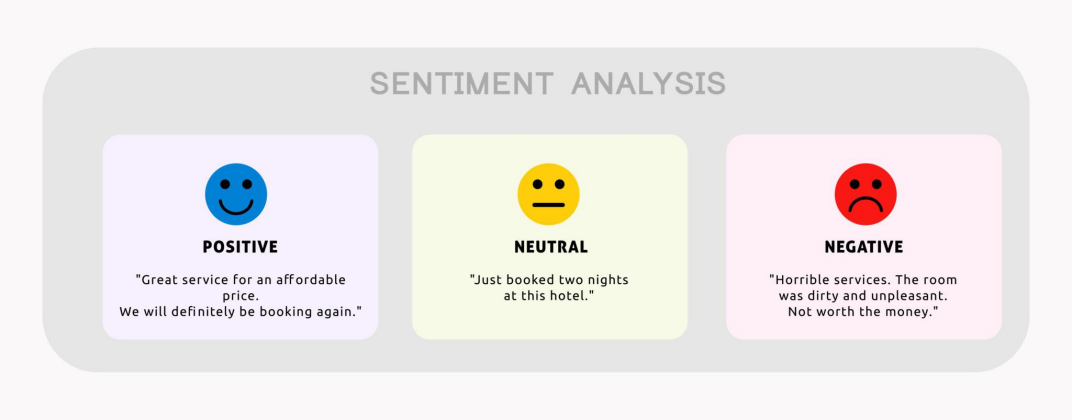


Tareas en procesamiento de lenguaje natural

- Clasificación de texto: Dada una secuencia de palabras, asignarle una clase a la secuencia.
- Detección de entidades nombradas: Detectar menciones de conceptos de interés.
- Transcripción de voz: Dado un discurso de voz, transcribir el texto de ese discurso
- Síntesis de voz: Dada una secuencia de palabras, generar un audio de la narración humana de ese texto.
- Síntesis de texto: Escribir creativamente una secuencia de palabras.
- Síntesis de imágenes: Dada una descripción en lenguaje natural, generar la imagen asociada.
- Traducción automática: Transformar secuencias de palabras entre distintos idiomas.

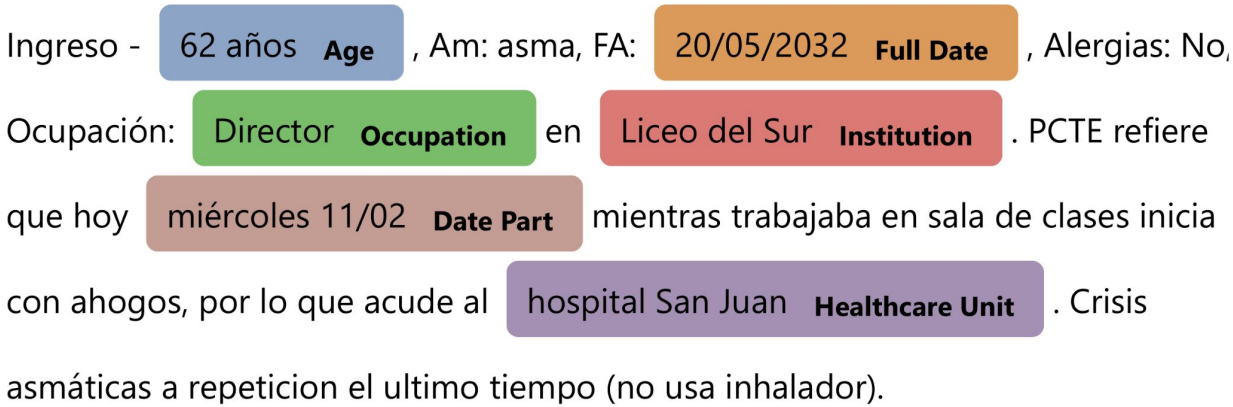
Clasificación de texto

La clasificación de texto es una técnica de aprendizaje automático que asigna un conjunto de categorías a una secuencia de palabras en la forma de texto libre no estructurado.



Detección de entidades nombradas

El reconocimiento de entidades nombradas es una subtarea de extracción de información que busca localizar entidades nombradas mencionadas en un texto libre no estructurado y clasificarlas dentro de un conjunto finito de categorías.



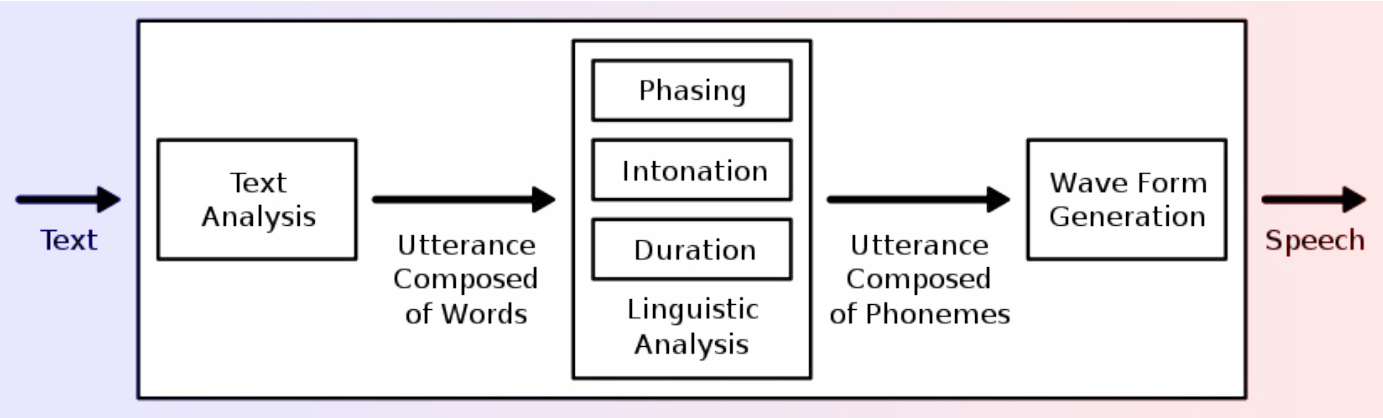
Transcripción de voz

El reconocimiento de voz es una área que desarrolla metodologías y tecnologías que permiten el reconocimiento y transcripción de lenguaje hablado hacía texto procesable por un computador.

Esta tarea se dedica a procesar un señal digital de audio, representarla y después decodificar la representación como una secuencia de palabras en lenguaje natural.

Síntesis de voz

Es parte del área del NLP que se dedica al reconocimiento del discurso humano en lenguaje natural en donde se busca reconstruir la señal de audio que generó una secuencia de palabras en lenguaje natural.



Modelos de lenguaje

Los modelos de lenguaje son funciones que le asignan una probabilidad a una secuencia de palabras.

Con estos modelos podemos tener la habilidad de generar texto que se parece mucho al lenguaje natural hablado por un humano.

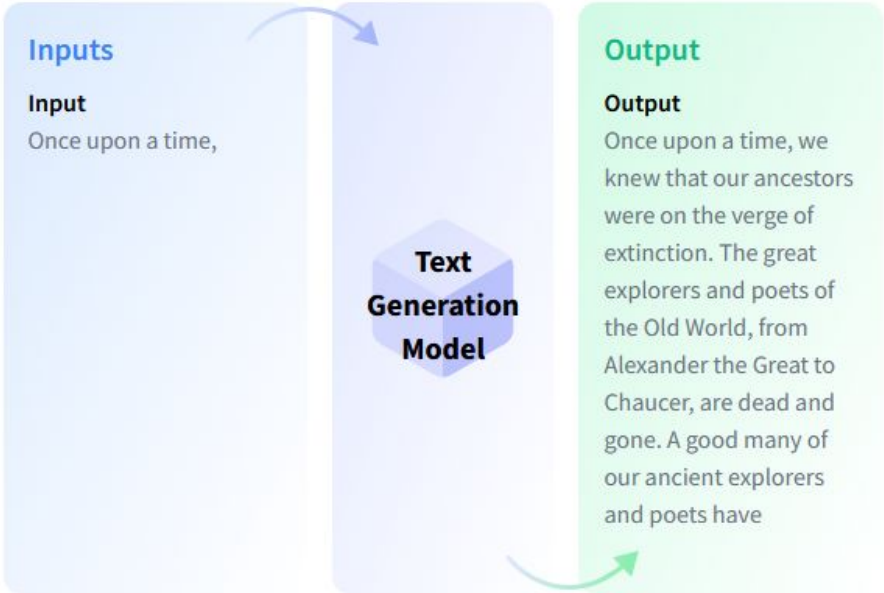
$$P(periodontitis, agresiva) > P(periodontitis, maligna)$$

$$P(agresiva|periodontitis) > P(maligna|periodontitis)$$

Síntesis de texto

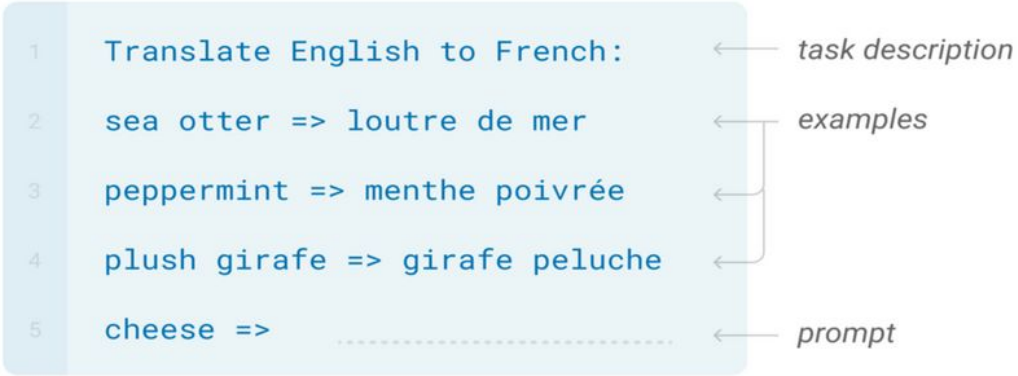
A través de un modelo de lenguaje podemos inferir la distribución de probabilidad de aparición de palabras y de una manera autoregresiva generar texto.

Típicamente en esta tarea se busca completar una secuencia de palabras que se le pasa al modelo.



Grandes modelos de lenguaje

A medida que los modelos de lenguaje basados en Transformers aumentan su cantidad de parámetros, los modelos tienen mejor rendimiento, pero llega un punto en que aparecen capacidades excepcionales, llamadas habilidades emergentes.



Síntesis de imágenes

Un modelo de síntesis de imágenes es un modelo de aprendizaje automático que toma una descripción en lenguaje natural y retorna una imagen que satisface la descripción pasada.

Se combina un modelo de lenguaje y un modelo generativo de imágenes para producir imágenes condicionadas por la representación de la descripción.

A photorealistic image of a person who knows too much natural language processing.

Traducción automática

La traducción automática convierte una secuencia de palabras desde un lenguaje de entrada hacia un lenguaje de salida.

Esta tarea se puede modelar como un problema secuencia a secuencia en donde se utiliza un modelo que representa el significado de una frase y otro modelo que reconstruye la representación en un idioma distinto.