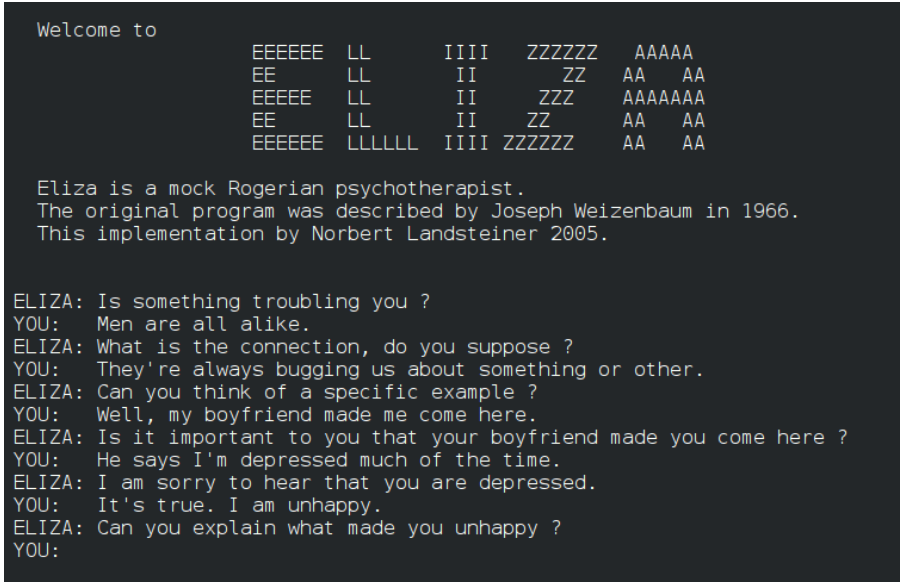
Procesamiento de Lenguaje Natural

Andrés Abeliuk, Fabián Villena

Introducción

- El NLP (Natural Language Processing) es una rama de la inteligencia artificial que se centra en la **interacción entre las computadoras y el lenguaje humano**.
- Su objetivo es permitir a las máquinas **comprender, interpretar y generar lenguaje humano de manera efectiva**.
- Utiliza técnicas de aprendizaje automático, lingüística computacional y otras disciplinas para procesar y analizar grandes cantidades de datos de texto.
- Se aplica en una variedad de campos, como la traducción automática, el análisis de sentimientos, la generación de texto, la extracción de información y más.



Eliza (Weizenbaum, 1966)

Desafíos en el NLP

El texto es un medio de **datos muy ambiguo y poco sistematizado**, lo que dificulta la extracción de información de estas fuentes.

La **variabilidad lingüística y la complejidad semántica** del lenguaje humano presentan desafíos únicos en el procesamiento automatizado del texto.

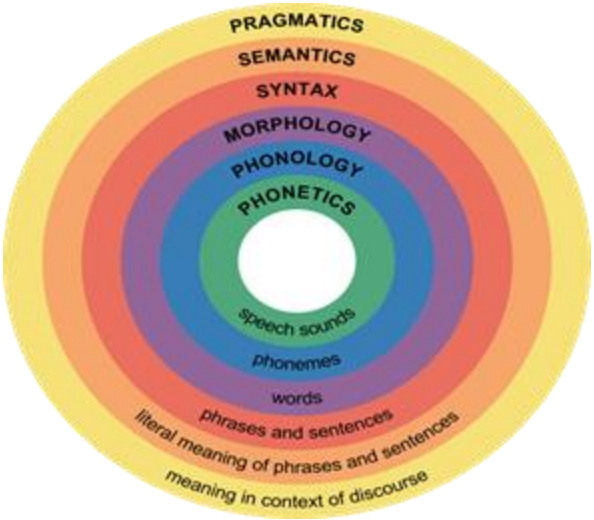
Existen **diferencias entre los datos de texto generados en distintos dominios** (por ejemplo, finanzas, medicina, redes sociales),

Un sistema diseñado para analizar texto en un dominio específico puede no funcionar adecuadamente en otro dominio.

La adaptación y la generalización de los modelos de NLP son desafíos importantes para garantizar su eficacia en una variedad de contextos y aplicaciones.

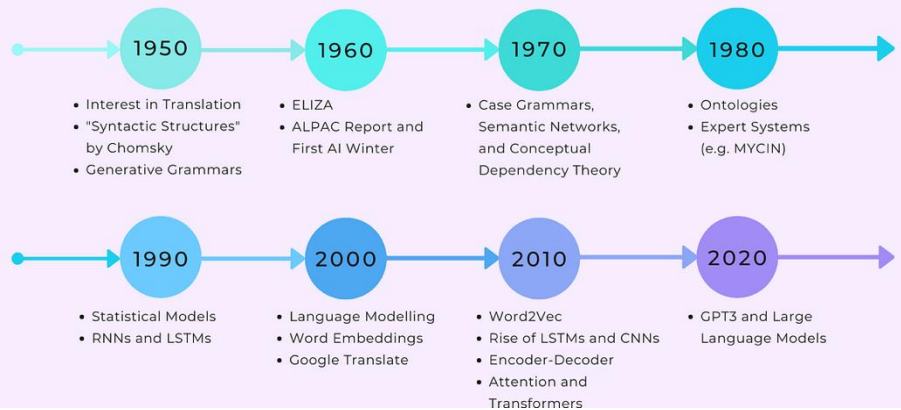
Niveles de análisis en NLP

Fonología: "casa" vs "caza" → sonidos diferentes
Morfología: "correr", "corredor", "corriendo" → misma raíz
Sintaxis: "María ama a Pedro" vs "Pedro ama a María" → orden importa
Semántica: "banco" = asiento vs institución financiera
Pragmática: "¿Podrías pasarme la sal?" → petición, no pregunta



Fonética	Fonología	Morfología	Sintaxis	Semántica	Pragmática
Proporciona la base para el reconocimiento y síntesis de voz.	Ayuda a entender los patrones de sonido en el lenguaje.	Crucial para el análisis de palabras y su descomposición en morfemas.	Esencial para entender la estructura de las oraciones.	Permite entender el significado de las oraciones y el contexto.	Ayuda a entender cómo el contexto afecta el significado.
Ejemplo: Asistentes virtuales como Siri y Alexa.	Ejemplo: Sistemas de texto a voz (TTS).	Ejemplo: Lematización y stemming en motores de búsqueda.	Ejemplo: Parsing y generación de texto.	Ejemplo: Clasificación de sentimientos y análisis de opiniones.	Ejemplo: Chatbots y sistemas de diálogo.

Técnicas de análisis de texto



Métodos Tradicionales

Basados en Reglas

- Reglas gramaticales y lingüísticas diseñadas manualmente.

- Ejemplo: Extracción de información mediante patrones predefinidos.

Métodos Modernos

Aprendizaje Automático

- Utilización de algoritmos para aprender patrones y relaciones en los datos.

- Ejemplo: Modelos de NLP que mejoran el rendimiento en diversas tareas.

Aprendizaje Automático en NLP

- Ventajas

- Capacidad para manejar datos complejos y variados.
- Mejora continua del rendimiento con más datos.

- Desafíos

- Necesidad de grandes volúmenes de datos para el entrenamiento.
- Explicabilidad limitada de las decisiones del modelo.

Complementariedad de Métodos

- Combinar métodos basados en reglas y aprendizaje automático.

- Ejemplo: Usar reglas lingüísticas para preprocesar texto antes de aplicar modelos de aprendizaje automático.

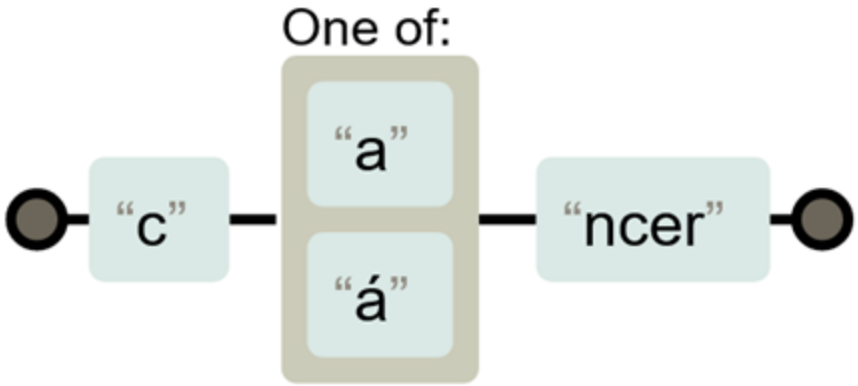
Métodos basados en reglas

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

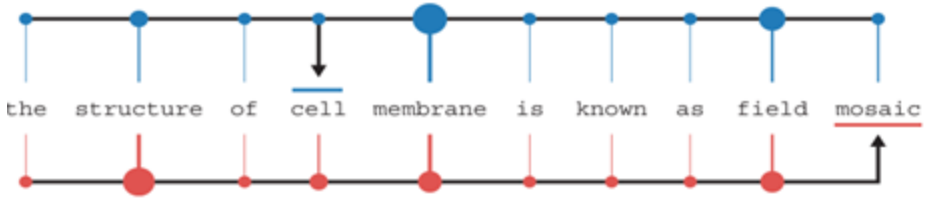
Lo más utilizado son las expresiones regulares para encontrar patrones dentro de cadenas de texto.

Ejemplos:
Extracción de Entidades Nombradas, Análisis de Sentimientos, Corrección Automática.

```
c[aá]ncer
```



Deep Learning



- A través de la utilización de redes neuronales artificiales somos capaces de modelar el lenguaje de una manera muy precisa.
- El aprendizaje profundo permite que el modelo aprenda automáticamente las características relevantes del texto a partir de los datos de entrenamiento.
- Estas características aprendidas pueden ser utilizadas para realizar tareas específicas en NLP,
    - como la clasificación de texto, el análisis de sentimientos y la generación de lenguaje natural.
- Con el advenimiento de nuevas arquitecturas basadas en autoatención llamadas Transformers han aparecido sorprendentes sistemas de NLP.

Generación de recursos lingüísticos

- Para entrenar un sistema de NLP, **se requieren muestras de ejemplo de lenguaje natural.**
    - Esto implica obtener y enriquecer grandes volúmenes de texto para ajustar un modelo de NLP.
- La generación de recursos lingüísticos por parte de humanos expertos es un proceso altamente costoso.
- El desarrollo de recursos lingüísticos de alta calidad es crucial para el desarrollo y el rendimiento efectivo de los sistemas de NLP en una amplia variedad de aplicaciones y contextos.

Tareas en procesamiento de lenguaje natural

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.

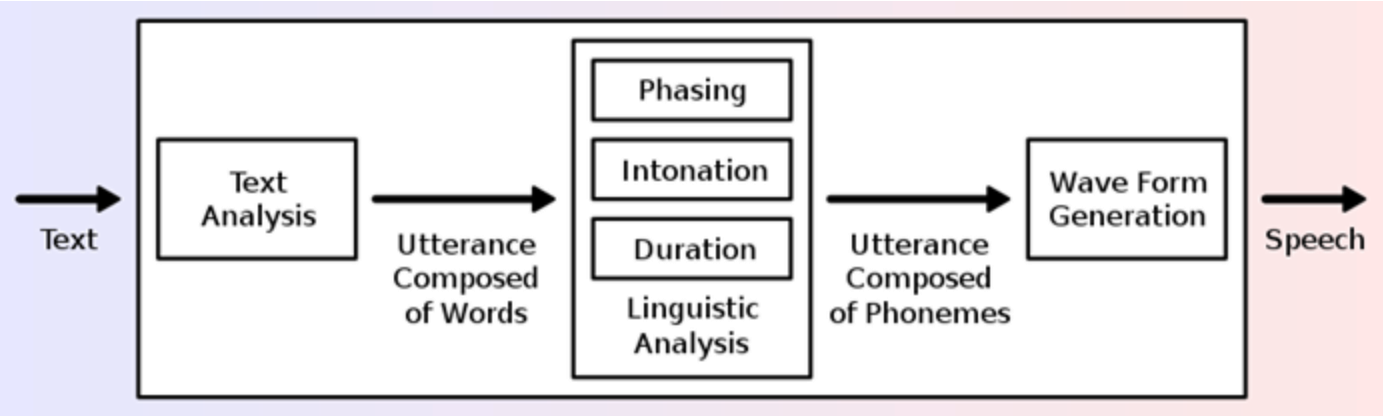
Transcripción de voz

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.



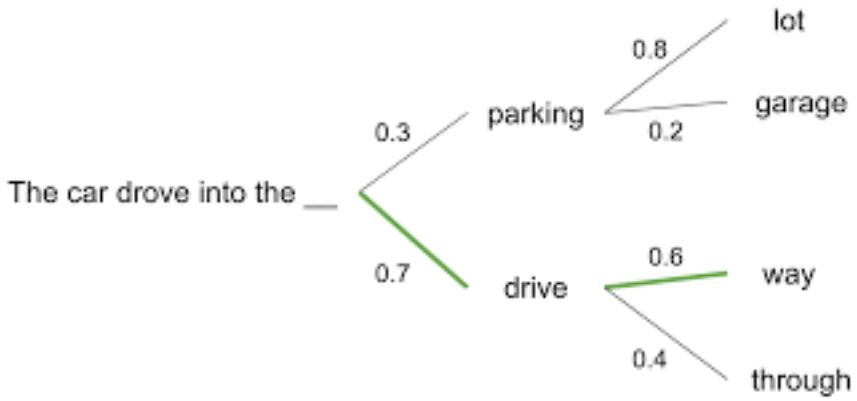
Modelos de lenguaje

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

Estos modelos son entrenados para predecir la probabilidad de ocurrencia de una palabra dada su historia o contexto en 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.

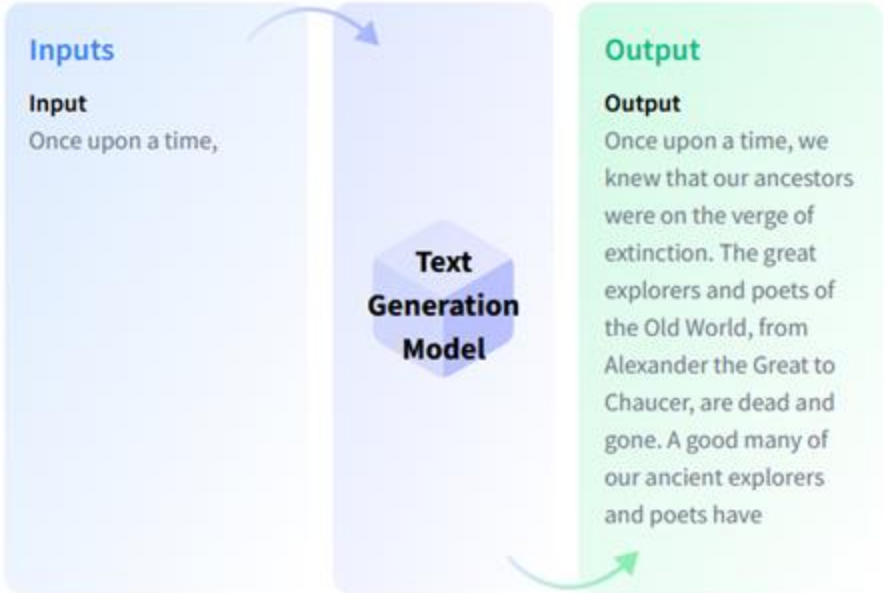
$$P(w_n|w_1,\ldots,w_{n-1})$$



Síntesis de texto

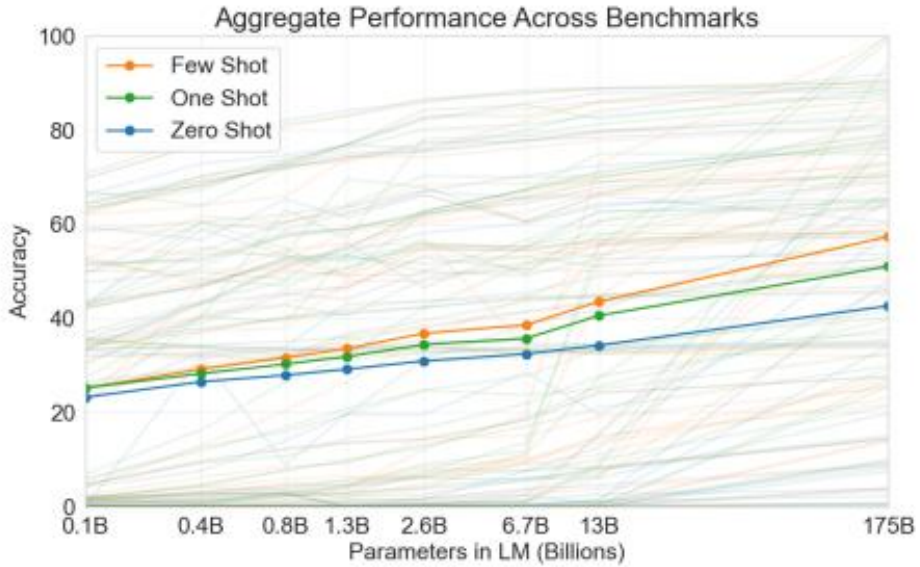
A través de un modelo de lenguaje podemos inferir la distribución de probabilidad de aparición de palabras y de una manera autoregresiva generar texto.

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.

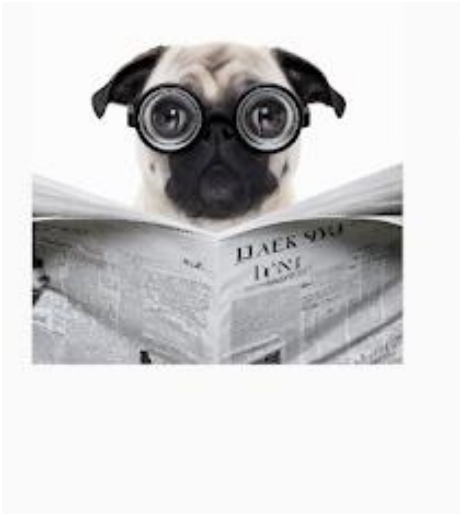


Fuente: Language Models are Few-Shot Learners

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.

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

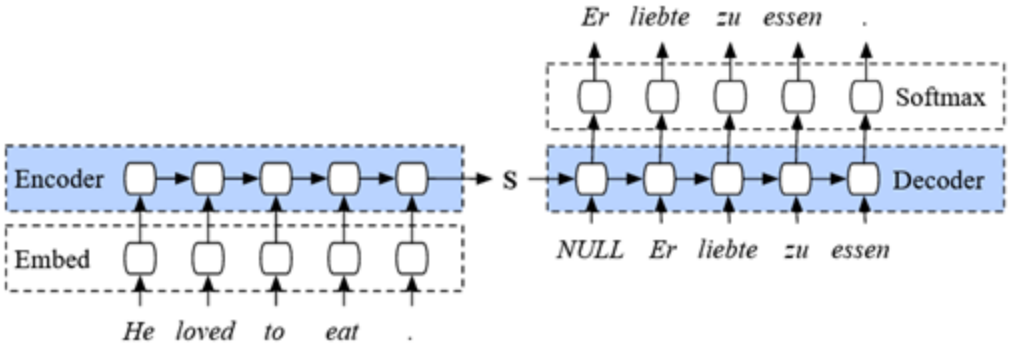


A dog reading the newspaper

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.



Ética y Sesgos en NLP

Sesgos en Modelos de Lenguaje

- Sesgo de género: "El doctor" vs "La enfermera"
- Sesgo racial y cultural en traducciones y clasificaciones
- Sesgos socioeconómicos en datasets de entrenamiento

Desinformación y Deepfakes

- Generación automática de noticias falsas
- Manipulación de opinión pública con bots

Privacidad y Datos Personales

- Análisis de conversaciones privadas y correos
- Extracción de información sensible sin consentimiento

Impacto Laboral

- Automatización de trabajos de escritura y traducción
- Necesidad de reentrenamiento y adaptación profesional

Contenido Curso

1. **Preprocesamiento de Texto y Exploración:**
   - Técnicas de preprocesamiento como tokenización y eliminación de stopwords.
2. **Semántica Vectorial:**
   - Representaciones numéricas básicas del texto mediante Bag of Words y TF-IDF.
3. **Word Embeddings:**
   - Uso de representaciones densas (Word2Vec, GloVe, FastText) para capturar relaciones semánticas entre palabras
4. **Tareas Supervisadas en NLP:**
   - Modelos de clasificación y reconocimiento de entidades (NER)
5. **Deep Learning y Redes Recurrentes:**
   1. Aplicación de redes neuronales recurrentes para modelar secuencias de texto.
6. **Transformers:**
   1. Arquitectura y aplicaciones de modelos Transformer en NLP.
7. **Grandes Modelos de Lenguaje y Habilidades Emergentes:**
   1. Estudio de modelos como GPT y BERT y sus capacidades emergentes.
8. **Cuantización y Fine-tuning Eficiente:**
   1. Optimización de grandes modelos de lenguaje para hacerlos más eficientes en la inferencia

Evaluación de Curso

- 2 tareas grupales

- Cada clase se descompone en:
  - clase expositiva
  - tutorial
  - Laboratorio