UT4

Procesamiento del Lenguaje Natural

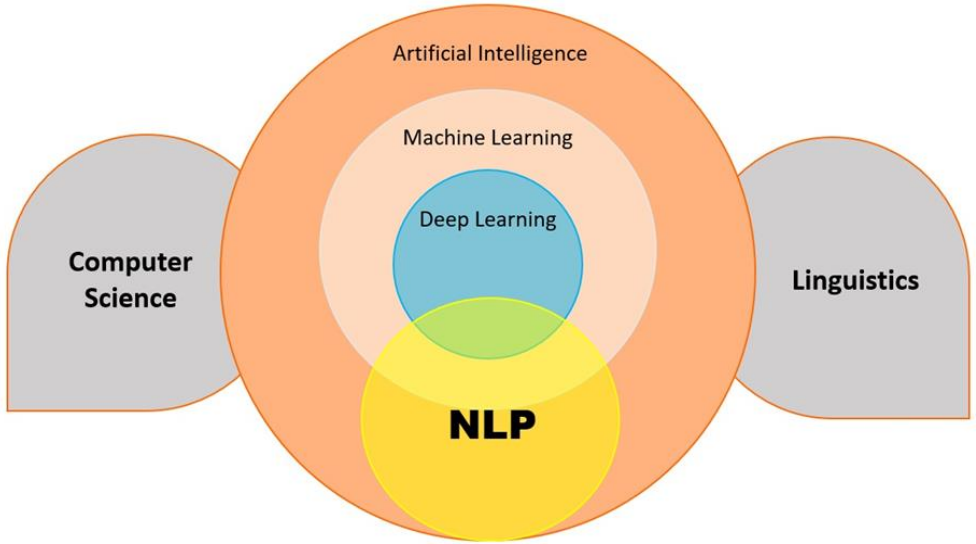
Fundamentos del Aprendizaje Automático

Profesor: Ing. Juan Francisco Kurucz

juan.kuruczsosa@ucu.edu.uy

¿Qué es NLP/PLN?

- NLP es una rama de la inteligencia artificial centrada en que las máquinas comprendan, procesen y generen lenguaje humano.
- Abarca tanto lenguaje escrito como hablado.
- Dos grandes enfoques:
  - NLU (Natural Language Understanding):
    - interpretar intención y significado.
  - NLG (Natural Language Generation):
    - producir texto coherente.
- Ejemplos: clasificar correos, traducir textos, conversar con usuarios, resumir documentos.

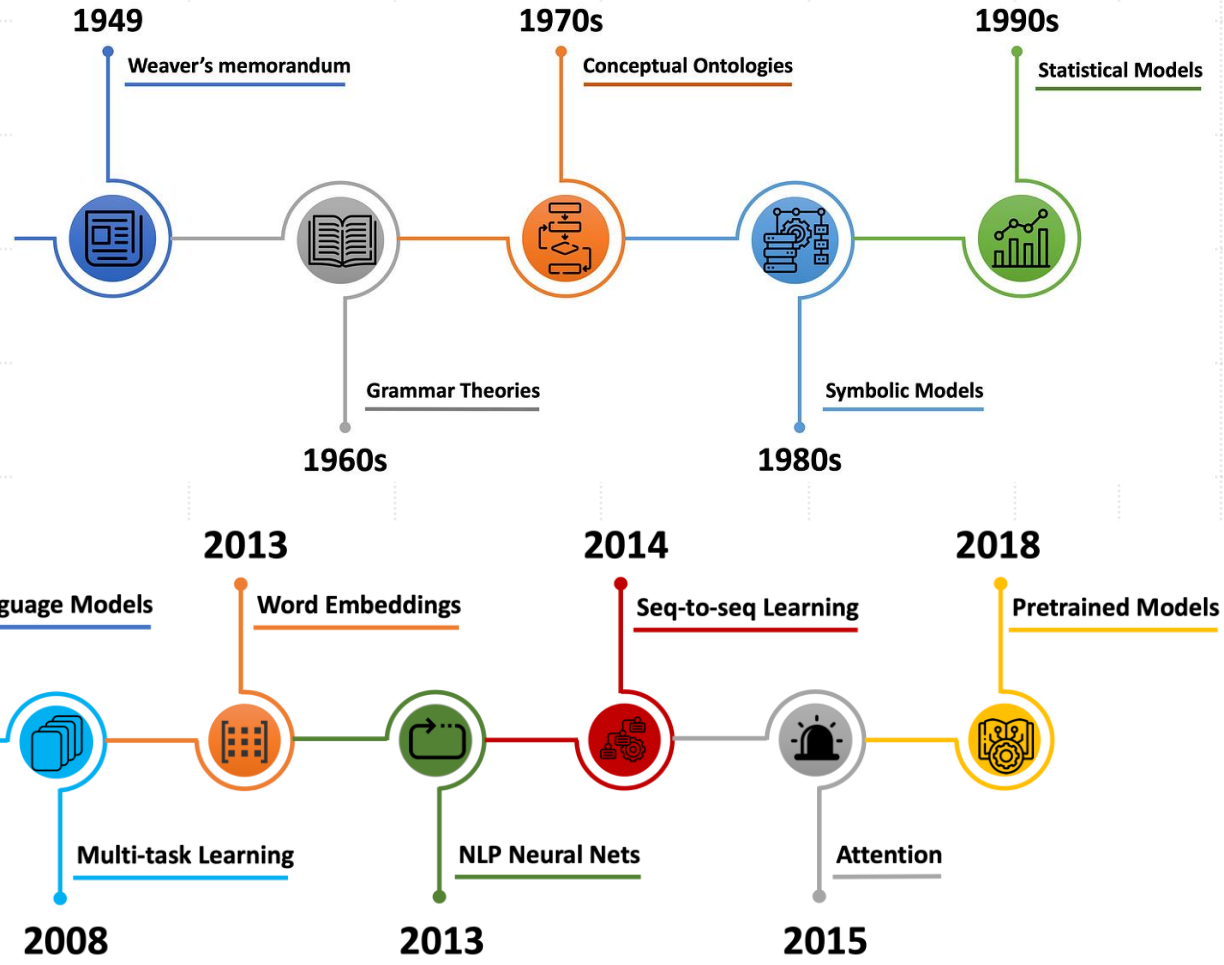


Aplicaciones clave hoy

- **Asistentes virtuales** (Siri, Alexa): interpretan comandos hablados y responden.
- **Moderación de contenido**: filtran mensajes ofensivos o spam.
- **Motores de búsqueda**: entienden intenciones y preguntas.
- **Chatbots en empresas**: automatizan soporte al cliente.
- **Resumen y traducción automática**: facilitan el acceso a información.

Evolución del NLP

- 1950s: sistemas por reglas.
- 1990s: estadística y corpus.
- 2000s: redes neuronales (RNN).
- 2013–15: word2vec, LSTM, seq2seq.
- 2017: Transformers.
- 2020+: LLMs masivos.

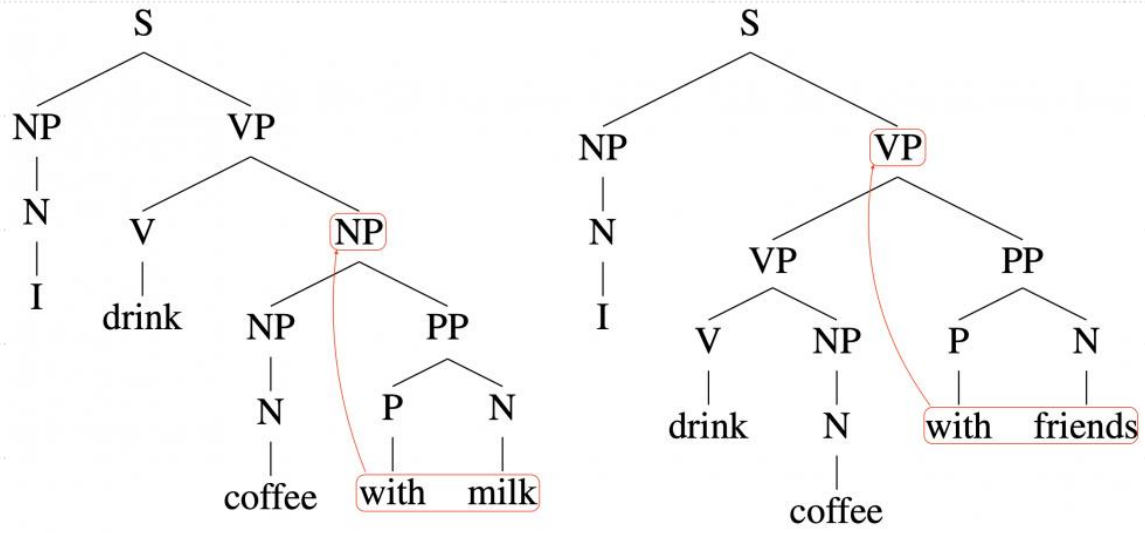


Orígenes (1950s-70s)

- Traducción automática temprana: Georgetown-IBM (1954).
- ELIZA (1966): imitaba a un terapeuta, basada en patrones de texto.
- Sistemas construidos a mano, sensibles a errores.
- Sin aprendizaje: reglas explícitas definidas por humanos.

Regla vs datos (70s-80s)

- Uso de gramáticas formales (Chomsky) y parsers sintácticos.
- Basados en conocimiento lingüístico experto.
- Problemas con ambigüedad y escalabilidad.
- Aporte: fundación formal y estructuras del lenguaje.

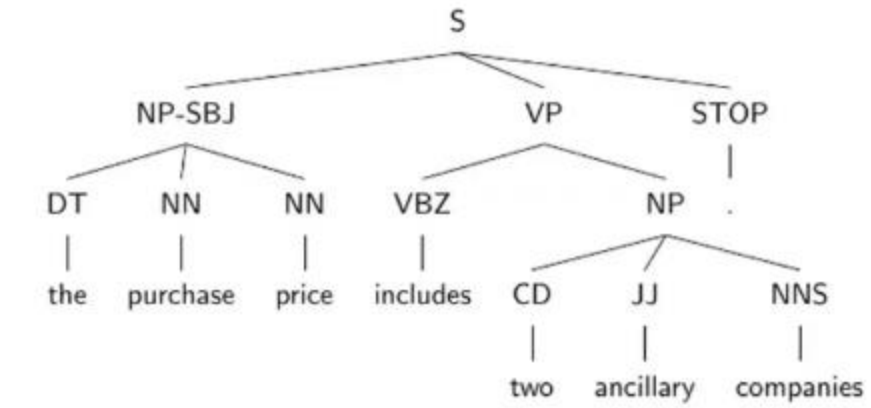


El giro estadístico (90s)

- Machine learning empieza a reemplazar reglas fijas.
- Modelos como HMM, Naive Bayes y n-gramas.
- Introducción de corpus como Penn Treebank.
- Se mide rendimiento con métricas como F1, Accuracy.

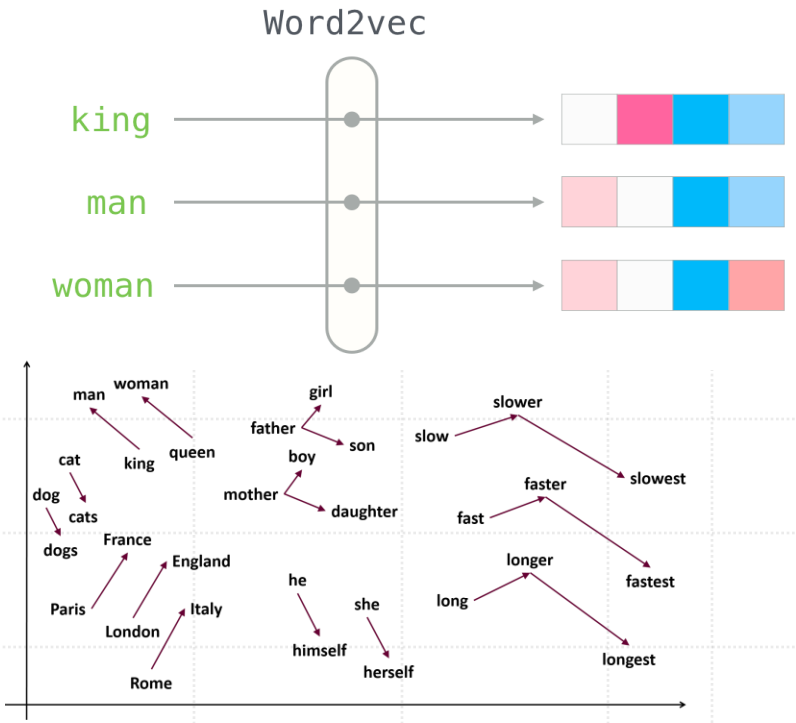
text = "The Margherita pizza is not bad taste"

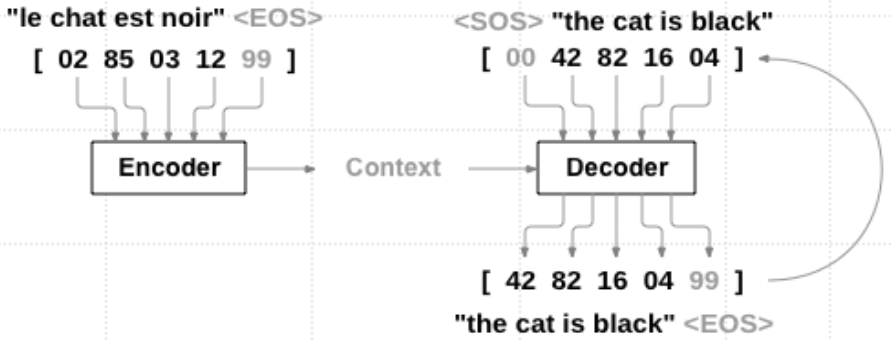
1-Gram	2-Gram	3-Gram
The	The Margherita	The Margherita pizza
Margherita	Margherita pizza	Margherita pizza is
pizza	pizza is	pizza is not
is	is not	is not bad
not	not bad	not bad taste
bad	bad taste	
taste		



Llega el deep learning (2000s)

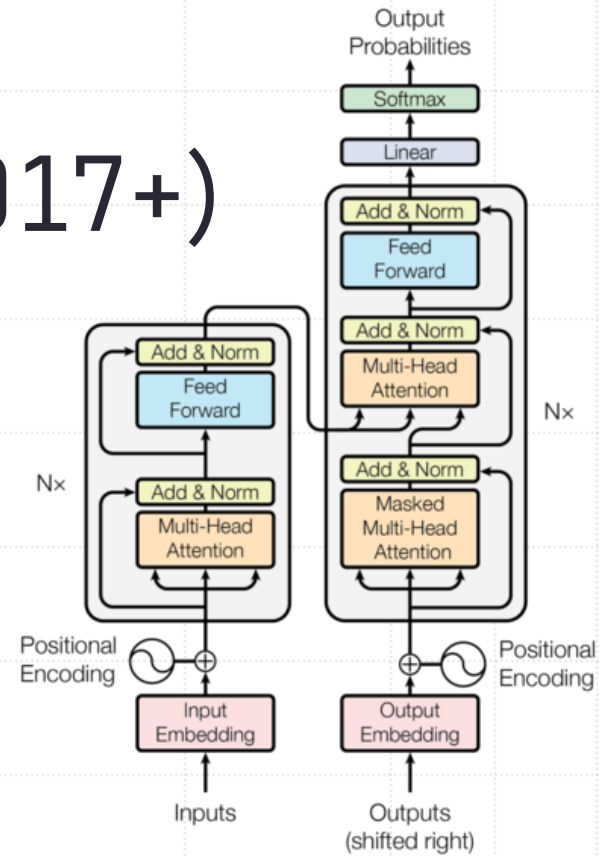
- word2vec y GloVe: embeddings densos que capturan contexto.
- RNN/LSTM: pueden recordar estados anteriores.
- seq2seq (encoder-decoder): revoluciona traducción y resumen.
- Limite: secuencialidad = entrenamiento lento y problemas con dependencias largas.

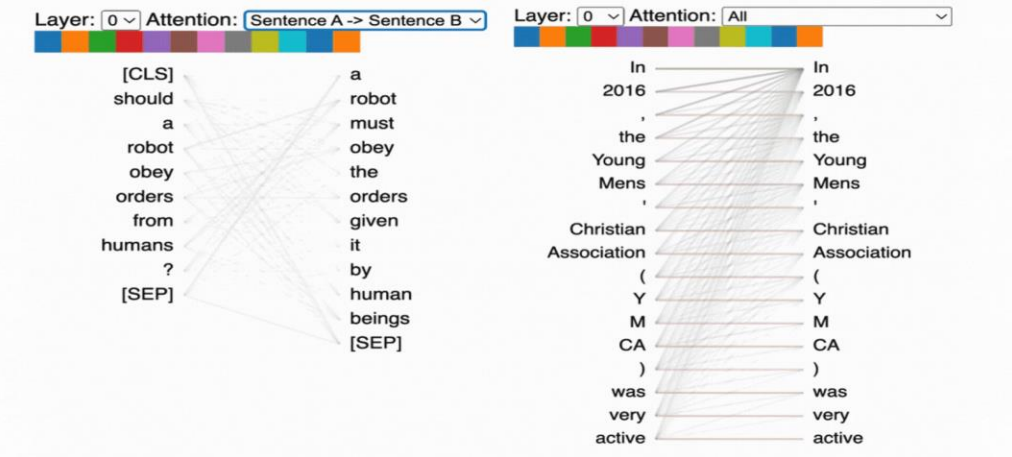




El salto: atención y Transformers (2017+)

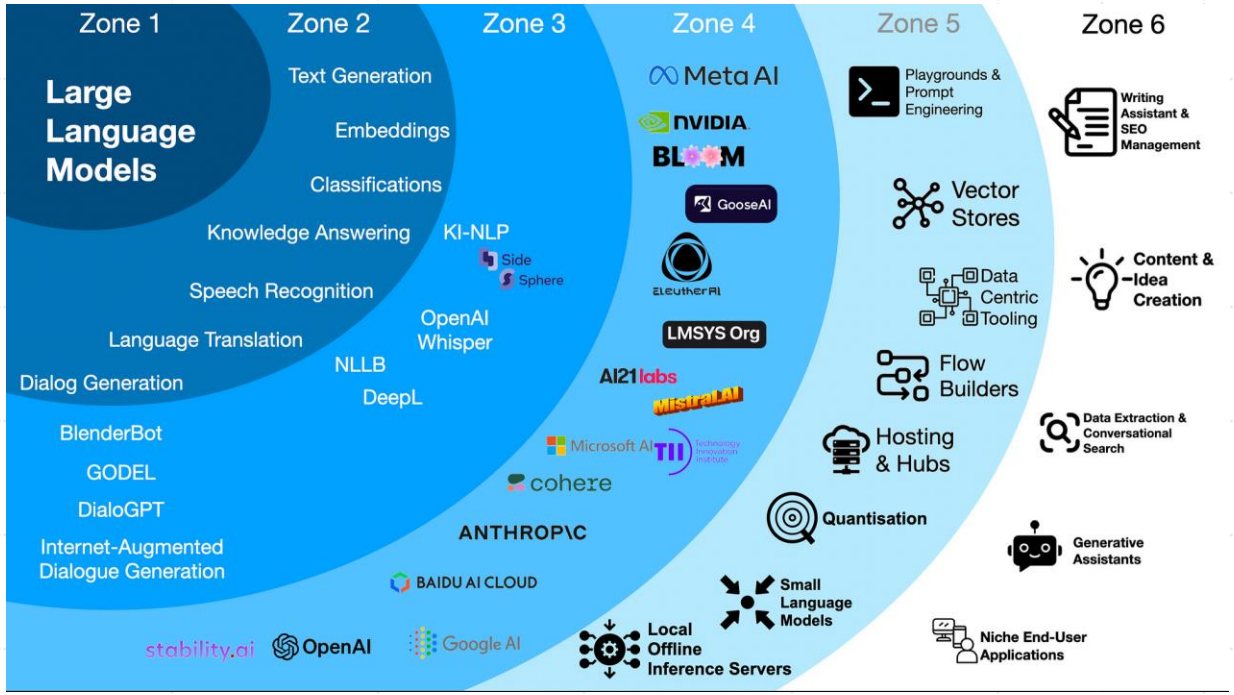
- Mecanismo de atención: permite enfocar en partes relevantes del input.
- Self-attention: cada token "mira" a todos los demás.
- Paraleliza entrenamiento, mejora contexto largo.
- Origen de BERT, GPT, T5 y más.





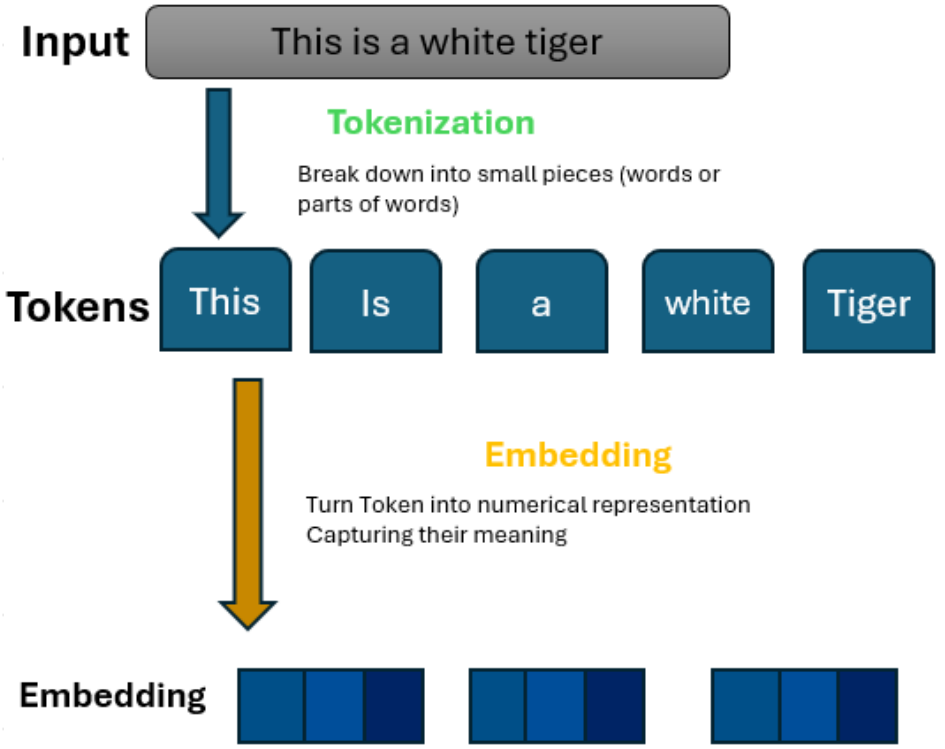
LLMs y su impacto

- GPT-3 (2020): 175B parámetros, zero-shot y few-shot learning.
- LLMs generalistas entrenados en Internet: generan, traducen, responden.
- Impacto: herramientas de productividad, educación, programación.
- Preocupaciones: alucinaciones, sesgos, dependencia tecnológica.



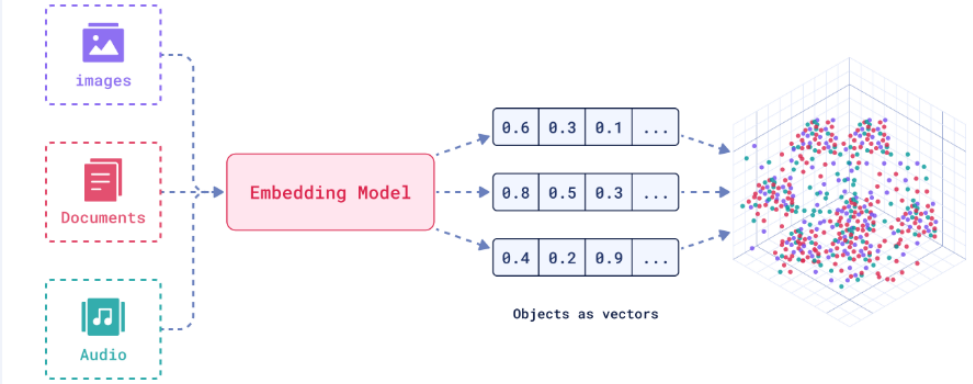
Tokens: la unidad de texto

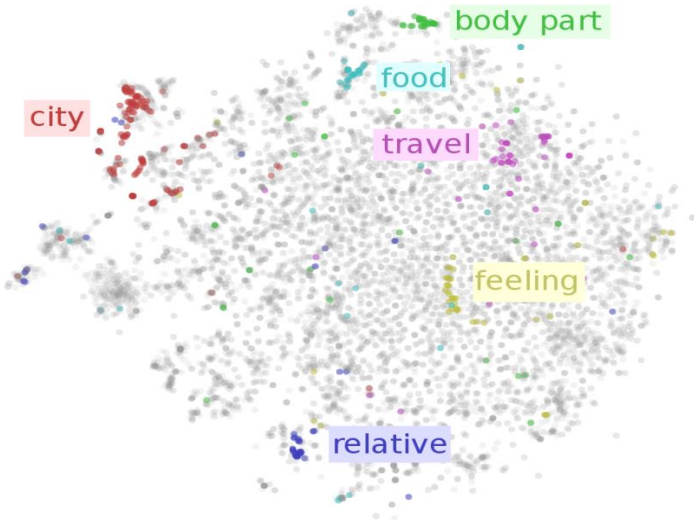
- NLP moderno opera sobre tokens, no palabras.
- Ej: "jugando" → ["jug", "ando"] con BPE.
- Tamaño del vocabulario se optimiza usando sub-palabras.
- Imprescindible para trabajar con modelos preentrenados.



Embeddings: representar significado

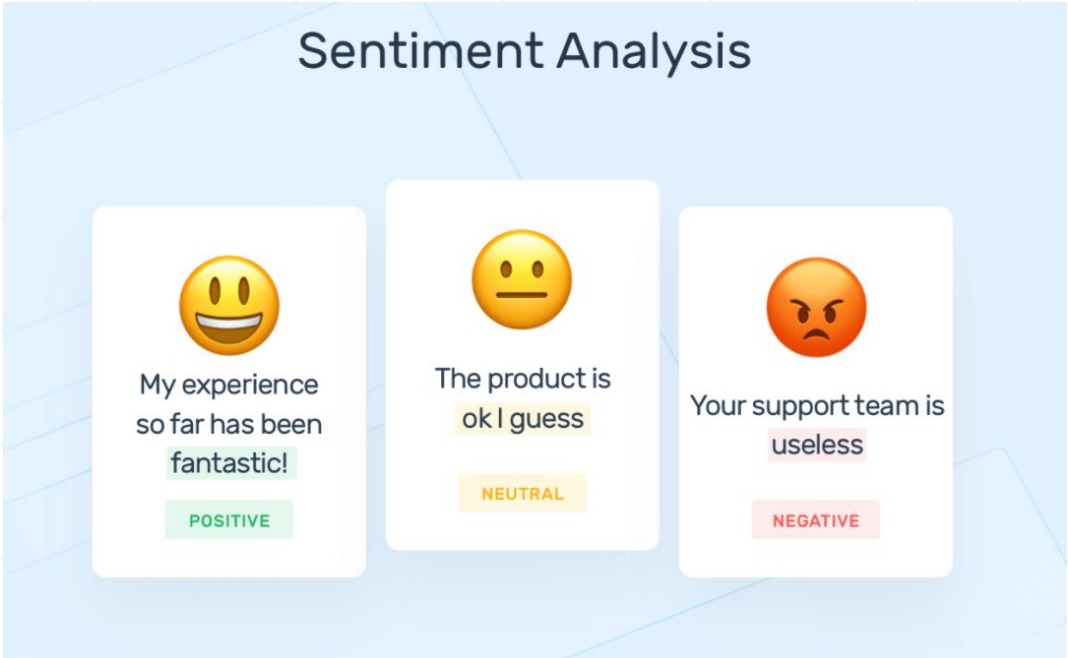
- Palabras y tokens se representan como vectores densos.
- word2vec: relaciones semánticas captadas en el espacio vectorial.
- Contextuales (BERT, GPT): el vector cambia según contexto





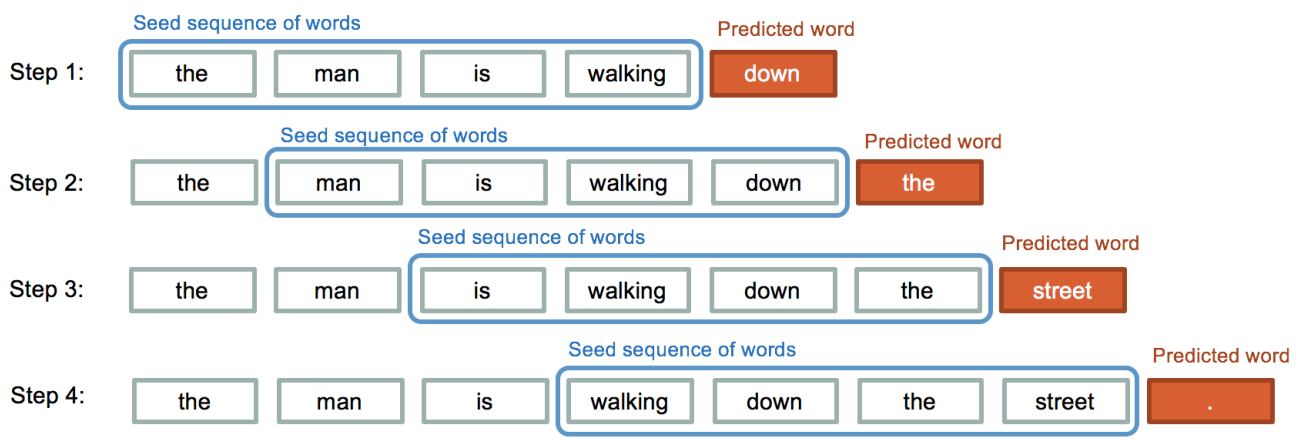
Clasificación de texto

- Tareas comunes: spam detection, sentimientos, temáticas.
- Input: texto; Output: categoría.
- En NLP moderno se usa fine-tuning de modelos como BERT.
- Ejemplo: "Esta película es genial" → positivo.



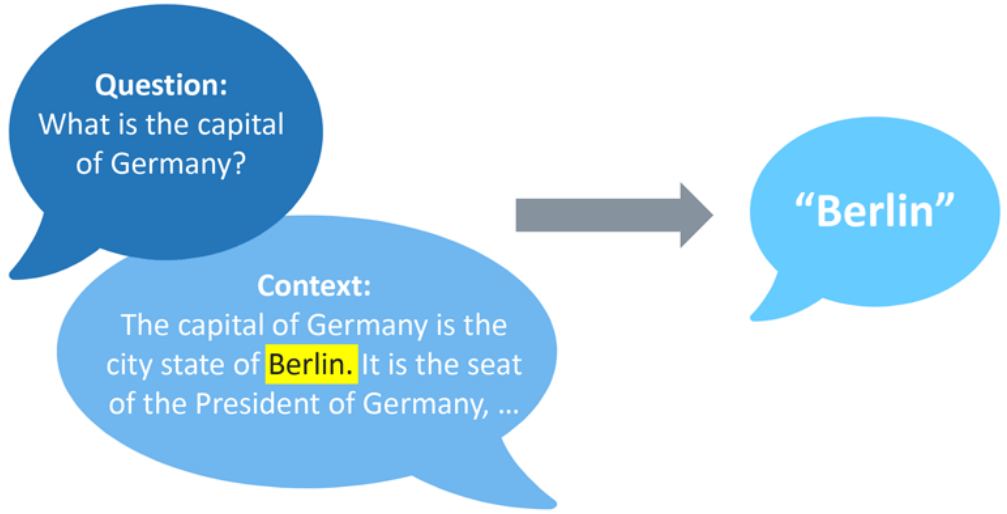
Generación de texto

- Predictivo: dado un contexto, generar lo que sigue.
- Ejemplo: "Buenos días, mi nombre es..."
- Modelos: GPT-2/3, T5, llama.
- Aplicaciones: resumen, autocompletado, traducción.



Pregunta-respuesta (QA)

- QA extractivo: busca respuesta dentro de un texto.
- QA generativo: redacta una respuesta nueva.
- QA abierto + Information Retrival: combinación con motores de búsqueda.
- Ejemplo: "¿Quién escribió Hamlet?" → Shakespeare.



Conversación

- Modelos de turno simple vs conversación continua (multi-turn).
- LLMs + memoria = coherencia en interacciones largas.
- LangGraph, RAG chatbots.
- Casos: soporte técnico, tutores virtuales.