Tareas supervisadas de PLN

Fabián Villena

Introducción

El aprendizaje supervisado es la base de muchas aplicaciones actuales del PLN. En esta clase exploraremos cómo los modelos aprenden a resolver tareas a partir de datos etiquetados y qué tipos de problemas se pueden abordar con este enfoque.

Aprendizaje automático

El aprendizaje automático permite a los sistemas aprender patrones a partir de datos, en lugar de ser programados manualmente. En PLN, esto significa que los modelos descubren regularidades del lenguaje a partir de ejemplos reales de uso.

- Un modelo aprende una función que asocia datos de entrada con una salida deseada.
- Se entrena con un conjunto de datos y se evalúa con datos nuevos.

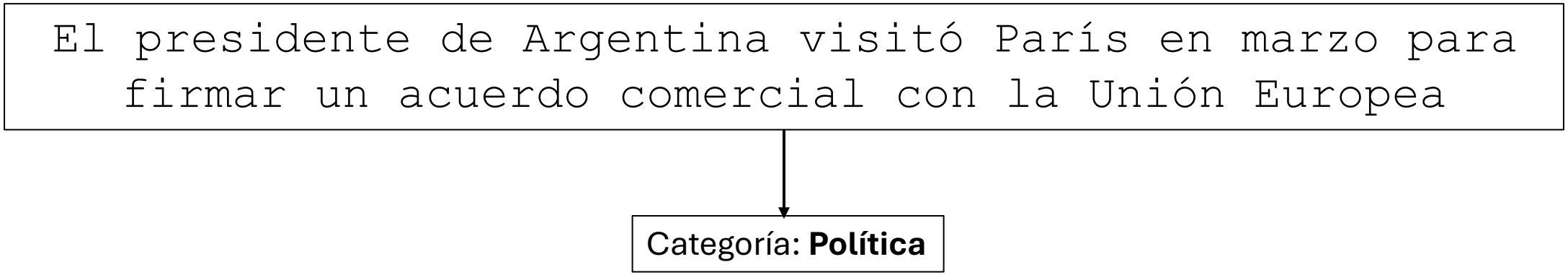
Aprendizaje supervisado

En el aprendizaje supervisado, el modelo aprende a partir de ejemplos etiquetados: para cada entrada conocida se dispone de la salida correcta. El objetivo es generalizar, es decir, predecir correctamente las etiquetas de nuevos ejemplos no vistos.

Necesitamos datos etiquetados representados de manera vectorial

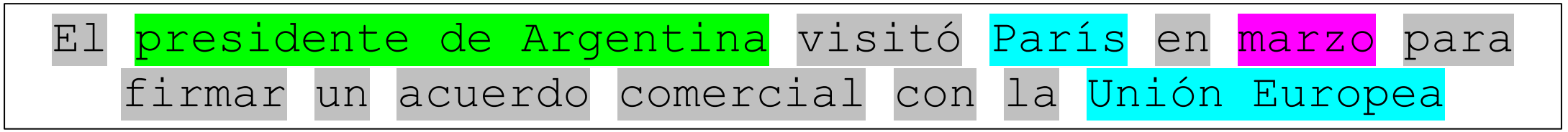
Clasificación de documentos

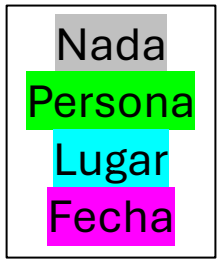
La clasificación de documentos consiste en asignar una o varias categorías predefinidas a un texto completo. Es una de las tareas supervisadas más comunes en PLN y se aplica tanto con modelos tradicionales como con *deep learning*.



Detección de entidades nombradas

La detección de entidades nombradas busca identificar y clasificar menciones de entidades dentro del texto, asignándoles una etiqueta semántica como persona, organización, lugar o fecha a cada uno de los tokens.





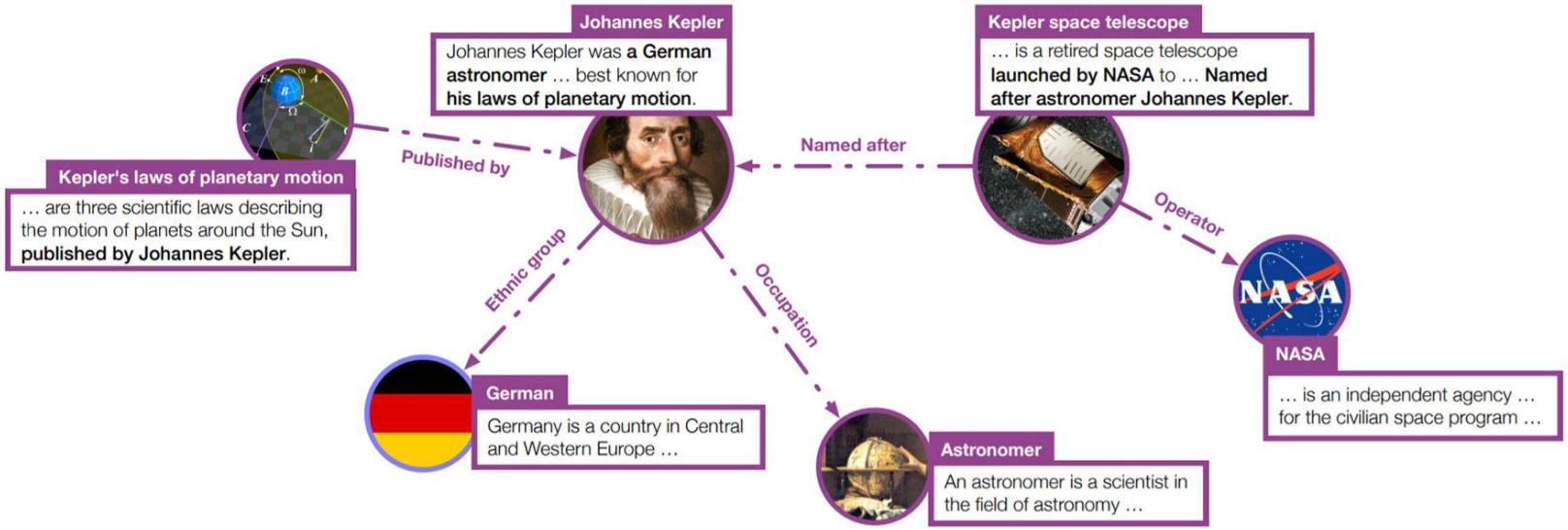
Enlazado de entidades

El enlazado de entidades (*entity linking*) consiste en asociar cada entidad detectada en el texto con una entrada única en una base de conocimiento, eliminando ambigüedades y proporcionando contexto semántico.

Grafos de conocimiento

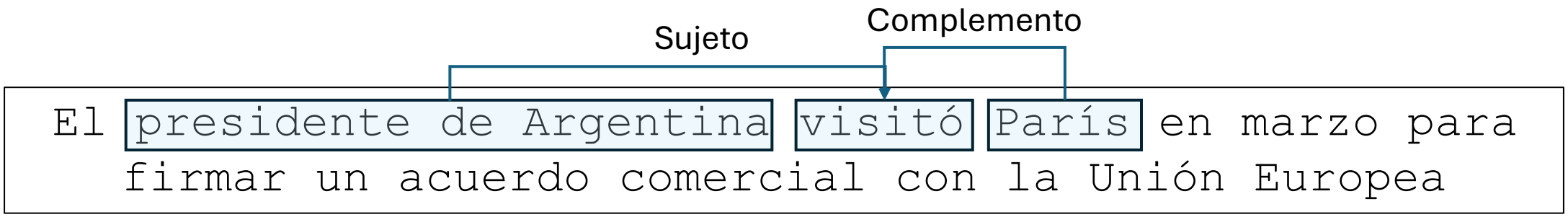
Un grafo de conocimiento representa información como un conjunto de entidades conectadas por relaciones semánticas.

Cada hecho se expresa como una tripleta (sujeto, predicado, objeto), permitiendo almacenar y razonar sobre conocimiento de manera estructurada.



Análisis de dependencia

El análisis de dependencia identifica la estructura sintáctica de una oración, representando cómo las palabras se relacionan entre sí mediante vínculos de dependencia (quién modifica o complementa a quién).



Extracción de información

La extracción de información busca identificar y estructurar hechos o relaciones significativas a partir de texto libre.

Convierte lenguaje natural en representaciones semiestructuradas útiles para bases de datos o grafos de conocimiento.

- Con los modelos anteriores se puede estructurar la información.

Modelamiento del lenguaje

El modelamiento del lenguaje consiste en aprender la probabilidad de aparición de las palabras en una secuencia, capturando regularidades y dependencias del lenguaje.

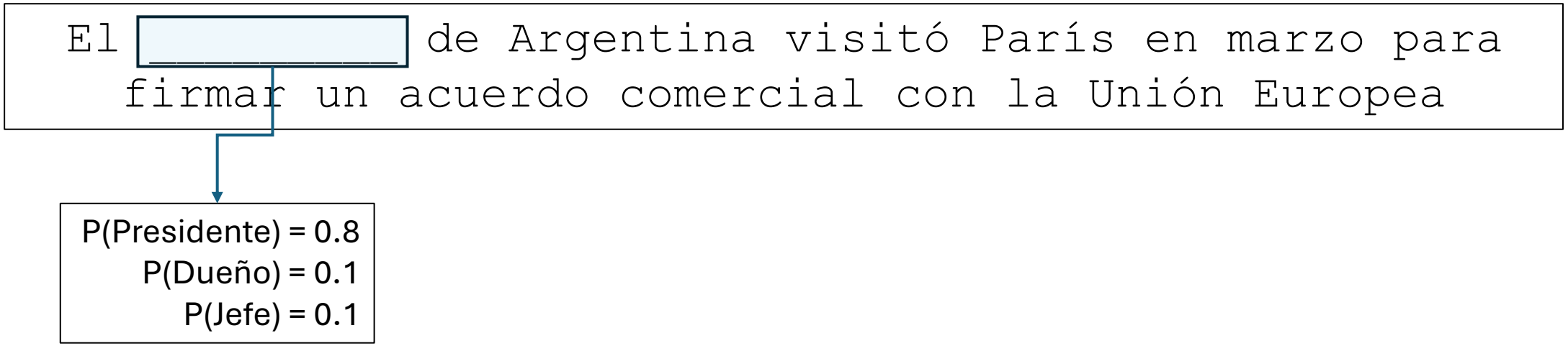
Es la base de muchas aplicaciones modernas de PLN, desde la predicción de texto hasta la generación y comprensión contextual.

$$P(\boxed{\texttt{El presidente de Argentina visitó París en marzo para firmar un acuerdo comercial con la Unión Europea}}) = 0.9$$

Modelos de lenguaje enmascarados

Los modelos de lenguaje enmascarados aprenden a predecir palabras ocultas dentro de una oración, comprendiendo el contexto bidireccional (izquierda y derecha).

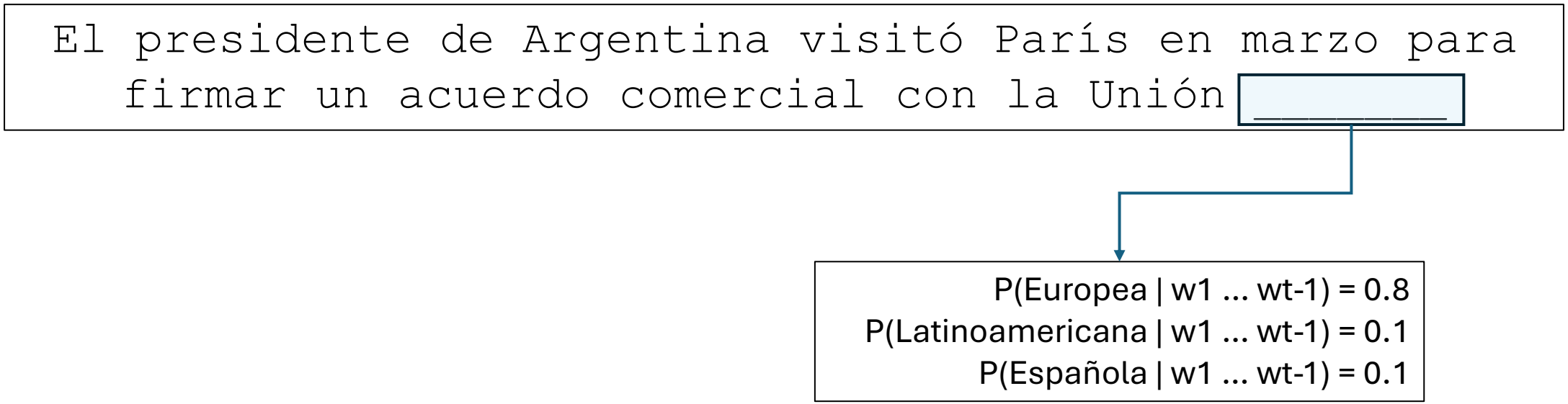
Este enfoque fue clave en la revolución de los modelos *transformer*.



Modelos de lenguaje causales

Los modelos de lenguaje causales aprenden a predecir la siguiente palabra en una secuencia, considerando únicamente el contexto previo.

Son la base de los modelos modernos de generación de texto.



Generación de texto

La generación de texto consiste en producir secuencias lingüísticamente correctas y semánticamente coherentes a partir de un contexto dado.

Aprovecha el conocimiento aprendido por los modelos de lenguaje para crear texto nuevo que imita el estilo y la estructura del lenguaje humano.

El presidente de Argentina visitó _
El presidente de Argentina visitó París _
El presidente de Argentina visitó París en _
El presidente de Argentina visitó París en Marzo _

Seguimiento de instrucciones

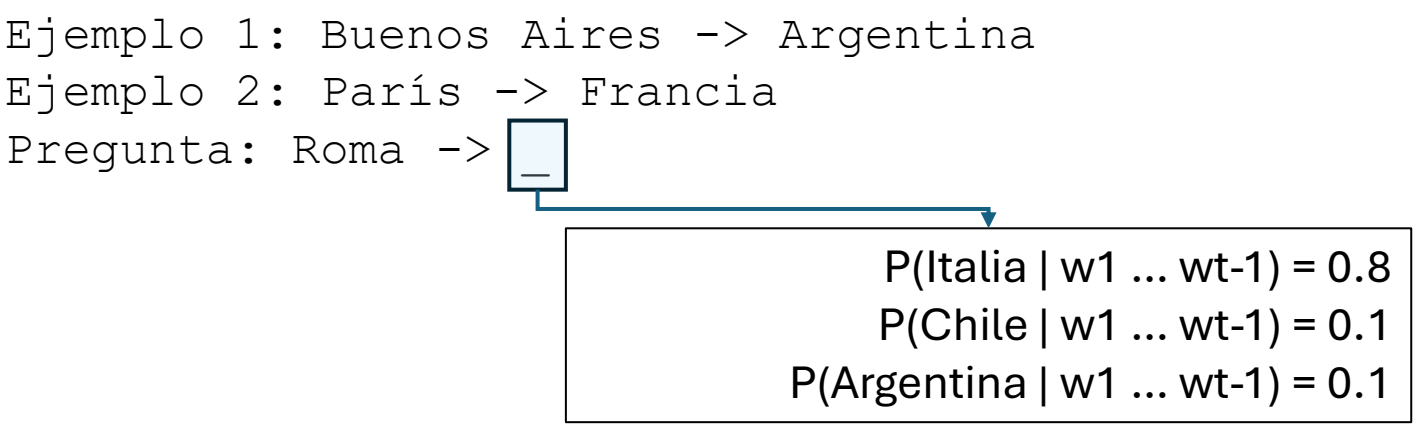
El seguimiento de instrucciones (instruction following) es la capacidad de un modelo de lenguaje para comprender una orden textual y generar una respuesta adecuada en función de esa instrucción.

Según el siguiente texto:
El presidente de Argentina visitó París en marzo para firmar un acuerdo comercial con la Unión Europea
¿Dónde se firmó el acuerdo?
Paris

Aprendizaje por contexto

El aprendizaje por contexto es la capacidad de un modelo de lenguaje para aprender una tarea nueva a partir de ejemplos incluidos directamente en el prompt, sin necesidad de reentrenamiento.

El modelo adapta su comportamiento utilizando solo la información contextual.



Traducción automática

La traducción automática convierte texto de un idioma a otro preservando su significado y estilo.

Es una de las tareas más emblemáticas del PLN y ha evolucionado desde reglas lingüísticas hasta modelos neuronales que aprenden directamente de grandes corpus bilingües.

Síntesis de voz

La síntesis de voz convierte texto escrito en audio hablado, permitiendo que los sistemas de PLN se comuniquen de manera oral y natural con las personas.

Combina procesamiento lingüístico, modelado acústico y generación de señal de audio.

```
El presidente de Argentina visitó París en marzo para
   firmar un acuerdo comercial con la Unión Europea
```

Reconocimiento automático del habla

El reconocimiento automático del habla convierte señales de audio en texto escrito, permitiendo que los sistemas comprendan el lenguaje hablado.

Es el complemento directo de la síntesis de voz.

El presidente de Argentina visitó París en marzo para firmar un acuerdo comercial con la Unión Europea

Síntesis de imágenes

La síntesis de imágenes consiste en generar imágenes a partir de descripciones textuales, combinando procesamiento del lenguaje natural y visión por computadora.

Permite crear contenido visual coherente con una instrucción o contexto lingüístico.

El presidente de Argentina visitó París en marzo para firmar un acuerdo comercial con la Unión Europea

Ciclo de aprendizaje supervisado en PLN