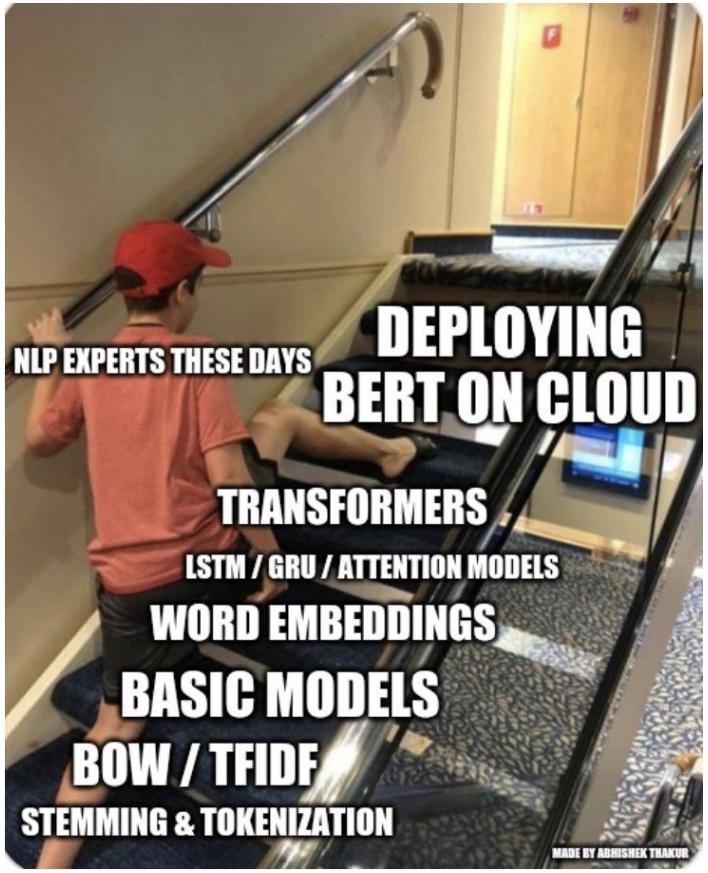
Introducción al procesamiento de texto y al lenguaje natural

Procesamiento de Lenguaje Natural Avanzado

Dra. Helena Gómez Adorno
helena.gomez@iimas.unam.mx

Bolsa de palabras → CNN → RNN/LSTM → Transformers

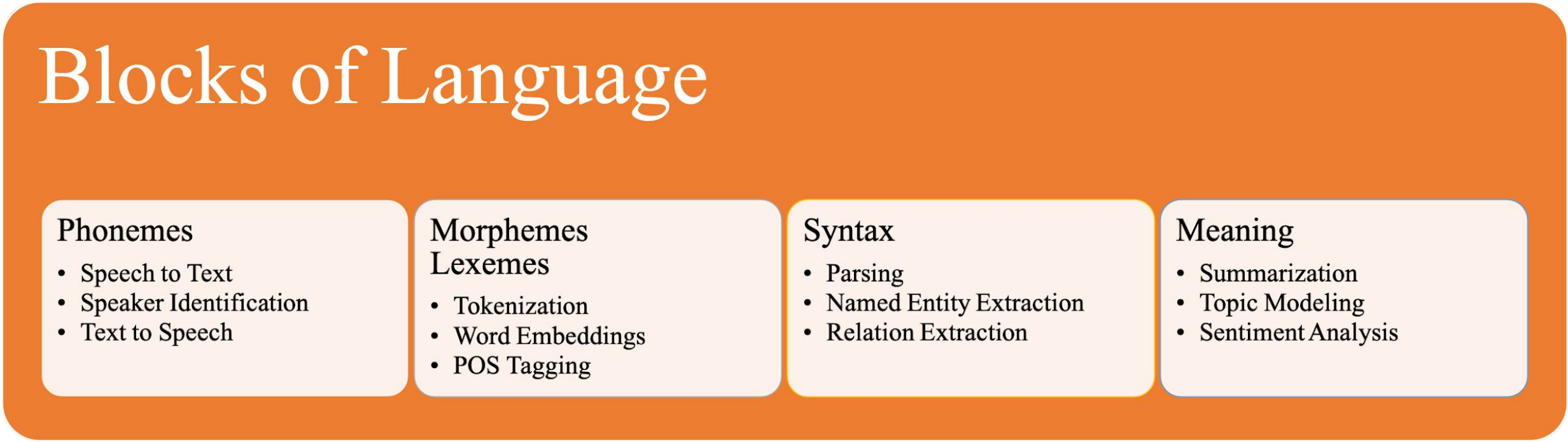


Aplicaciones de PLN

El procesamiento del lenguaje natural, o, más ampliamente hablando, la lingüística computacional es un campo muy activo en la lingüística aplicada porque está estrechamente relacionado con el desarrollo de la tecnología del lenguaje humano. El término **procesamiento del lenguaje natural** es un término más específico que se refiere al subcampo de la computación que se ocupa de los métodos para **analizar**, **modelar** y **comprender** el lenguaje humano.

Core	General App	Industry Specific
• Text Classification • Information Extraction • Conversational Agent • Information Retrieval • Question Answering Systems	• Spam Classification • Calendar Event Extraction • Personal Assistants • Search Engines • Jeopardy!	• Social Media Analysis • Retail Catalog Extraction • Health Records Analysis • Financial Analysis • Legal Entity Extraction

Tareas de PLN y dimensiones de análisis del lenguage



Desafíos de PLN

- **Ambigüedad**: El lenguaje natural es inherentemente ambiguo, con palabras y frases que a menudo tienen múltiples significados o interpretaciones dependiendo del contexto.
  - Golpeó el armario con el palo y **lo** rompió
  - Es tan bueno como Juan Perez.
  - El trofeo no cabe en el maletín marrón porque es demasiado [**pequeño/grande**]. ¿Qué es demasiado [**pequeño/grande**]?
- **Creatividad**: El lenguaje humano es altamente creativo, lo que permite infinitas posibilidades de expresión e interpretación. Esta creatividad plantea desafíos para los sistemas de PLN al generar y comprender patrones lingüísticos novedosos o no convencionales.

Desafíos de PLN

- **Diversidad**: El lenguaje exhibe una diversidad significativa en diferentes contextos, culturas, dialectos y hablantes individuales. Los sistemas de PLN entrenados en conjuntos de datos específicos pueden encontrar dificultades cuando se enfrentan a variaciones lingüísticas que difieren de los datos de entrenamiento.
- **Conocimiento común (Contexto)**: Comprender el lenguaje a menudo requiere conocimientos previos y contexto que pueden no estar explícitamente establecidos en el texto. Los sistemas de PLN deben ser capaces de incorporar conocimientos comunes e información contextual para interpretar y generar lenguaje con precisión.

Inteligencia artificial, aprendizaje automático, aprendizaje profundo y PLN

Inteligencia artificial

Es el campo de estudio

Aprendizaje automático

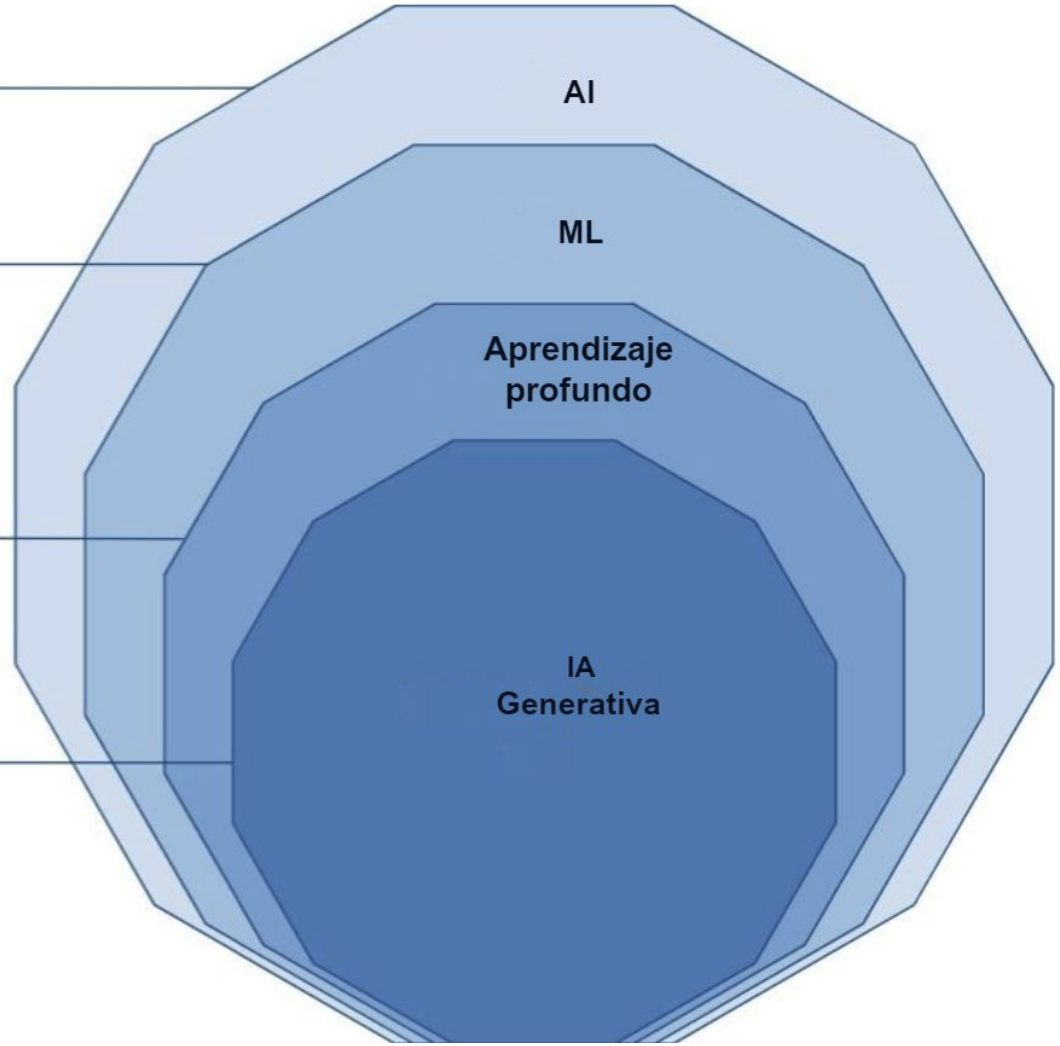
Es una rama de AI que se centra en la creación de Máquinas inteligentes que aprenden de los datos. Otra sucursal muy conocida dentro de AI es Mejoramiento.

Aprendizaje profundo

Es un subconjunto de métodos de aprendizaje automático, Basado en Redes Neuronales Artificiales.
Ejemplos: CNN, RNNS

AI generativo

Un tipo de RNA que generan datos que son similar a los datos con los que fue entrenado.
Ejemplos: GANS, LLMS



Enfoques de PLN

- **PLN basado en heurística**
- **PLN basado en aprendizaje automático (machine learning)**
- **PLN basado en aprendizaje profundo (deep learning)**

PLN basado en heurística

Se basan en reglas y patrones predefinidos para procesar el lenguaje natural. Estas reglas suelen ser elaboradas por lingüistas o expertos en el dominio basándose en principios y conocimientos lingüísticos. Los métodos heurísticos implican el diseño de algoritmos que codifican reglas lingüísticas para realizar tareas como el etiquetado de parte del discurso, el reconocimiento de entidades nombradas y el análisis sintáctico. Si bien puede ser efectiva para manejar fenómenos y tareas lingüísticas específicos, a menudo requiere un extenso esfuerzo manual para diseñar y mantener las reglas, y puede tener dificultades para manejar la complejidad y la variabilidad del lenguaje natural.

- **Ejemplos:**
- Análisis de sentimientos basado en diccionarios
- WordNet para relaciones léxicas
- Expresiones regulares
- **Fortalezas:**
- Las reglas basadas en el conocimiento específico del dominio pueden reducir de manera eficiente los errores que a veces son muy costosos

PLN basado aprendizaje automático (machine learning)

Aprovechan modelos estadísticos y algoritmos para aprender patrones y estructuras a partir de grandes cantidades de datos de texto anotados. Estos modelos están entrenados en conjuntos de datos etiquetados para extraer automáticamente características y hacer predicciones para varias tareas de PLN. Sobresalen en tareas como la clasificación de texto, el análisis de sentimientos y la traducción automática, y pueden manejar fenómenos lingüísticos más complejos en comparación con los enfoques basados en la heurística. Requieren cantidades significativas de datos etiquetados para el entrenamiento y pueden tener dificultades para manejar la ambigüedad y la variabilidad lingüística.

- **Tipos de aprendizaje automático:**
  - Supervisado vs. No supervisado
  - Clasificación vs regresión
- **Tres pasos comunes para el aprendizaje automático**
  - Extracción de características de textos
  - Usar la representación de características

  para entrenar un modelo
  - Evaluar y refinar el modelo

- **Métodos comunes:**
- Bayes ingenuo
- Regresión logística
- Máquina de vectores de soporte
- Modelo oculto de Markov
- Campo aleatorio condicional

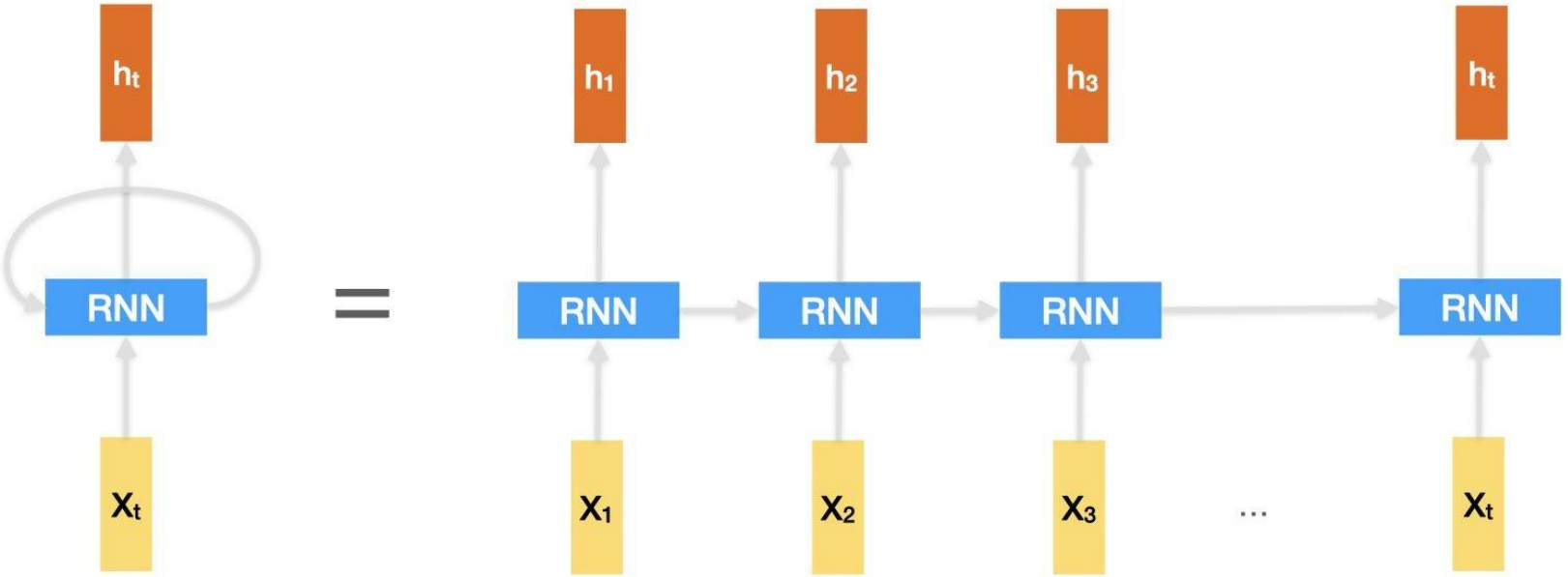
PLN basado aprendizaje automático (machine learning)

Implica el uso de arquitecturas de redes neuronales, particularmente redes neuronales profundas con múltiples capas, para modelar y procesar datos de lenguaje natural. Los modelos de aprendizaje profundo, como las redes neuronales recurrentes (RNN), las redes de memoria largo-corto plazo (LSTM) y los modelos de transformador, han demostrado un éxito notable en varias tareas de PLN debido a su capacidad para aprender representaciones jerárquicas de datos de texto. Estos modelos pueden aprender automáticamente patrones y dependencias intrincados en los datos, lo que los hace altamente efectivos para tareas como el modelado de lenguaje, la generación de secuencia a secuencia y la incrustación contextual de palabras. El aprendizaje profundo para PLN ha llevado a avances significativos en áreas como la traducción automática, el resumen de texto y la respuesta a preguntas, y sigue siendo un área activa de investigación en el campo.

- **Métodos comunes:**
  - Red neuronal convolucional (CNN)
  - Modelos de secuencia
    - Red neuronal recurrente (RNN)
    - Memoria a largo-corto plazo (LSTM)
  - Atención y Transformers

Fortalezas de los modelos de secuencia

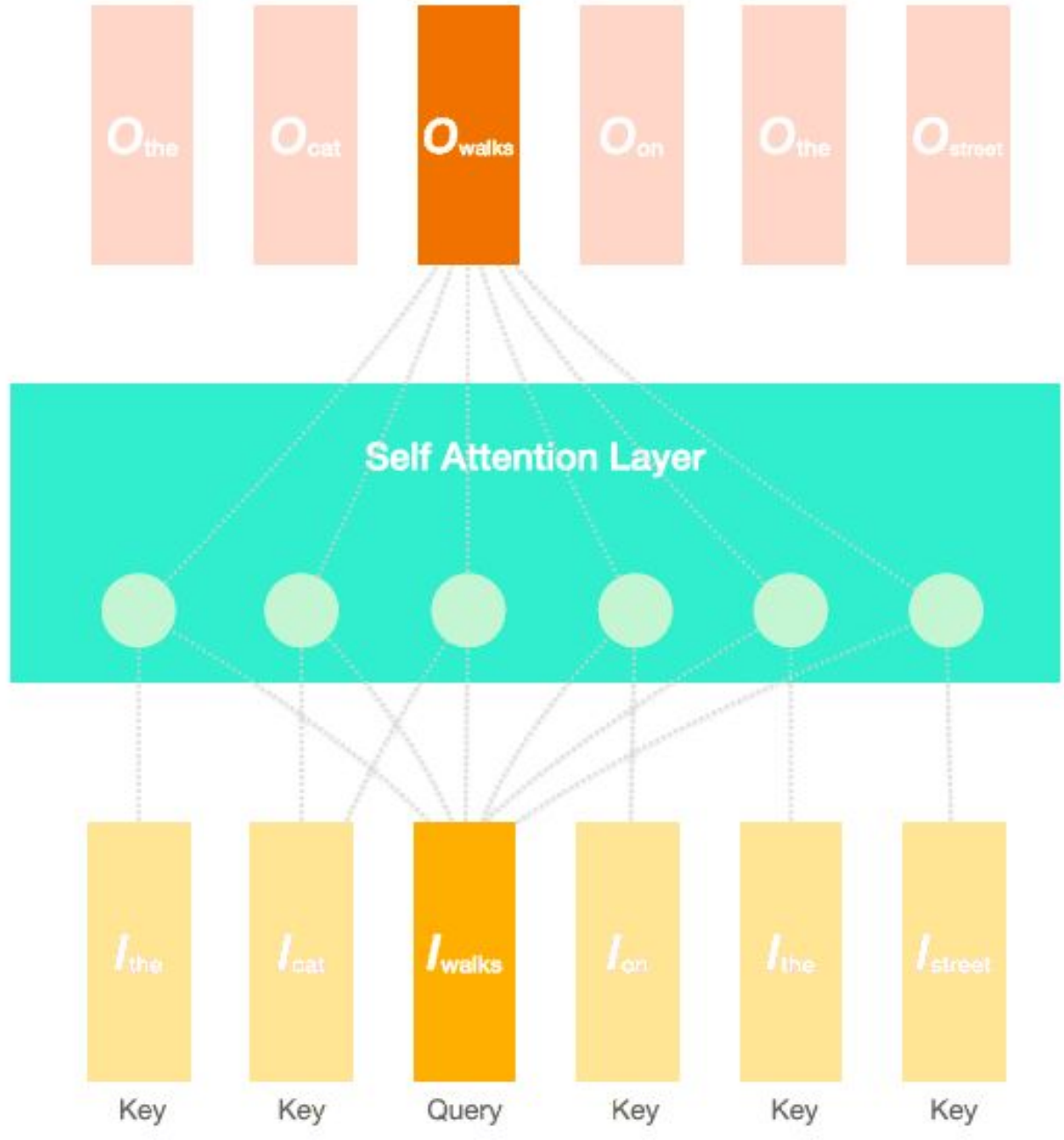
- Refleja el hecho de que una oración en el lenguaje fluye de una dirección a otra.
- El modelo puede leer progresivamente un texto de entrada de un extremo a otro.
- El modelo tiene unidades neuronales capaces de recordar lo que ha procesado hasta ahora.



Transformers

- Estos modelos sobresalen en la captura de dependencias de largo alcance y se han convertido en la columna vertebral de muchos sistemas de PLN de última generación.
- Aprovecha el poder de la **atención**. Es un mecanismo que permite al modelo aprender qué partes de la secuencia de entrada (es decir, la información contextual) son más relevantes para la tarea objetivo, dando más peso a las palabras o frases importantes.
- Ciertas partes de una historia son más importantes que otras. El mecanismo de atención en el aprendizaje profundo funciona como un foco, destacando estas partes importantes a medida que se desarrolla la historia. Ayuda al modelo a centrarse en la información más relevante en cada paso del procesamiento, al igual que cómo podría prestar atención a los detalles clave mientras lee.

Self-Attention



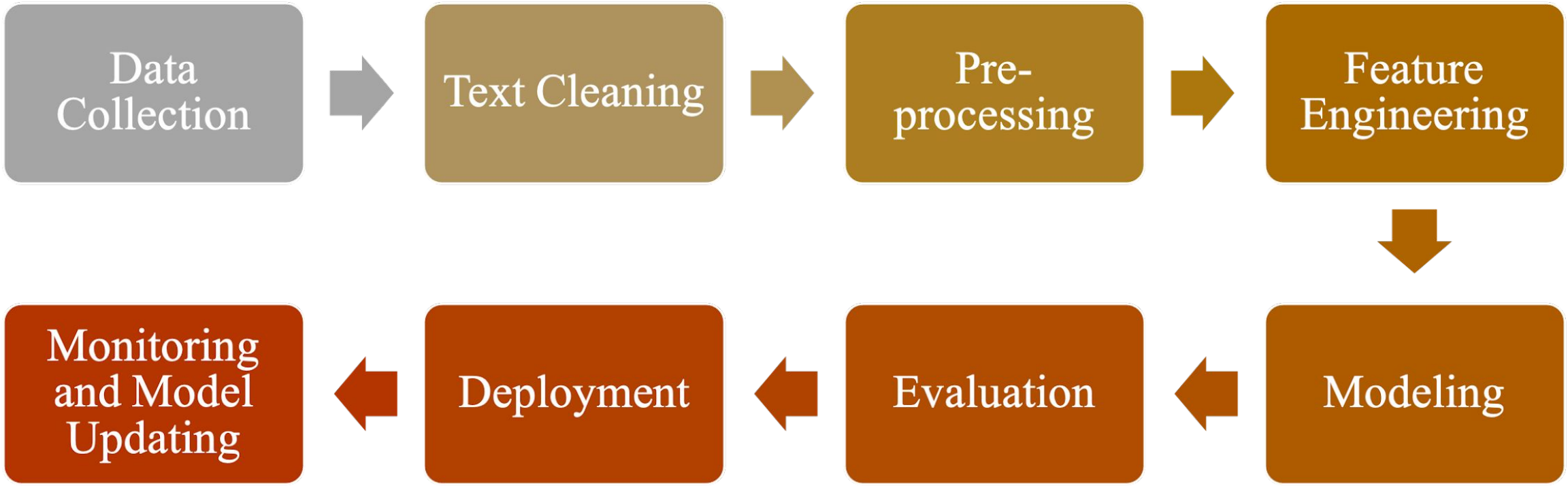
Transferir el aprendizaje

- El aprendizaje por transferencia es una técnica de aprendizaje automático en la que un modelo preentrenado en una tarea o conjunto de datos se reutiliza como punto de partida para entrenar un modelo en una tarea o conjunto de datos diferente pero relacionado.
- Esta reutilización a menudo se realiza a través del ajuste fino de los modelos preentrenados.
- En Deep Learning, el aprendizaje por transferencia es importante porque permite aprovechar los modelos preentrenados para lograr un buen rendimiento con datos limitados específicos de tareas.
- Al adaptar los modelos preentrenados a dominios o tareas específicas, el aprendizaje por transferencia reduce el tiempo de entrenamiento y mejora el rendimiento en los objetivos específicos de la tarea.

El aprendizaje profundo NO lo es todo

- **Sobreajuste en pequeños conjuntos de datos**
- **Adaptación de dominio (géneros)**
- **Modelos interpretables**
- **Alto costo del aprendizaje profundo**
- **Implementación en el dispositivo**

Pipeline general de PLN



Recopilación de datos

- **Configuración ideal**: Tenemos todo lo que necesitamos.
- Etiquetas y anotaciones
- Muy a menudo estamos lidiando con escenarios menos que ideales

Escenarios menos que ideales:

- Conjuntos de datos iniciales con anotaciones/etiquetas limitadas
- Conjuntos de datos iniciales etiquetados/extraídos basados en expresiones regulares o heurísticas
- Conjuntos de datos públicos (cf. Búsqueda de conjuntos de datos de Google o kaggle)
- Intervención del producto (La intervención del producto se refiere a la manipulación o curación deliberada de datos por parte de desarrolladores de productos o propietarios de plataformas, lo que puede introducir sesgos o distorsiones en el conjunto de datos).

Aumento de datos

El aumento de datos en el procesamiento del lenguaje natural es una técnica utilizada para aumentar la diversidad y la cantidad de datos de entrenamiento aprovechando las propiedades del lenguaje para crear nuevas muestras de texto que son sintácticamente similares a los datos de texto de origen originales.

- **Tipos de estrategias:**
- Reemplazo de sinónimos
- Reemplazo de palabras relacionadas (basado en métricas de asociación)
- Traducción de retroceso (backtranslation)
- Reemplazar entidades
- Agregar ruido a los datos (por ejemplo, errores ortográficos, palabras aleatorias)
- Paráfrasis automatica - Resumen automatico

Extracción y limpieza de texto

- Extracción de textos sin procesar de los datos de entrada
  - HTML
  - pdf
- Información relevante frente a información irrelevante
  - Información no textual
  - Marcado
  - Metadatos
- Formato de codificación

Extracción y limpieza de texto

Extraer textos de páginas web

- Extraer contenido textual del sitio web es una forma muy común de obtener datos. Requiere un estudio detallado de la estructura del contenido HTML de las páginas web.

Extracción de textos de PDF

- Tesseract es un motor de reconocimiento de texto (OCR) de código abierto, disponible bajo la licencia Apache 2.0. Se puede usar directamente, o (para los programadores) usando una API para extraer texto impreso de imágenes. Admite una amplia variedad de idiomas. Tienes que instalarlo **manualmente** antes de poder usarlo en Python.

Normalización Unicode: revisar la documentación de unicodedata para obtener más detalles sobre la normalización de caracteres.

Algunas notas sobre Unicode

Unicode: Unicode es un estándar para codificar caracteres en sistemas digitales. A cada carácter se le asigna un punto de código único, que es un valor numérico que representa ese carácter. Por ejemplo, el punto de código Unicode para el carácter "我" es U+6211

- ord(): In Python, the ord() function returns the Unicode code point (in decimal format) for a given character. For example, to get the Unicode code point for the character “我”, you would use ord("我"), which returns 25105.
- hex(): The hex() function in Python converts an integer to its hexadecimal representation. So, if we pass the Unicode code point of “我” to hex(), it will return its hexadecimal representation. For example, hex(25105) will return '0x6211', where '0x' indicates that the following digits are in hexadecimal format.

Limpieza (Preprocesamiento)

Preprocesamiento frecuente

- Eliminación de stop words
- Stemming y/o lematización
- Eliminación de dígitos/sintuaciones
- Normalización de casos

Preprocesamiento específico de la tarea

- Normalización de Unicode
- Detección de idioma
- Mezcla de código (codeswitching, ejemplo: spansglish, jopará (Guaraní-Español))

Preprocesamiento (Enriquecimiento)

Anotaciones automáticas

- Etiquetado POS
- Parsing sintáctico/dependencia
- Reconocimiento de entidades nombradas
- Resolución de coreferencia

Recordatorios importantes para el preprocesamiento

- No todos los pasos son necesarios
- Estos pasos NO son secuenciales
- Estos pasos dependen de la tarea y del idioma
- Objetivos
  - Normalización de texto
  - Tokenización de texto
  - Enriquecimiento/Anotación de texto

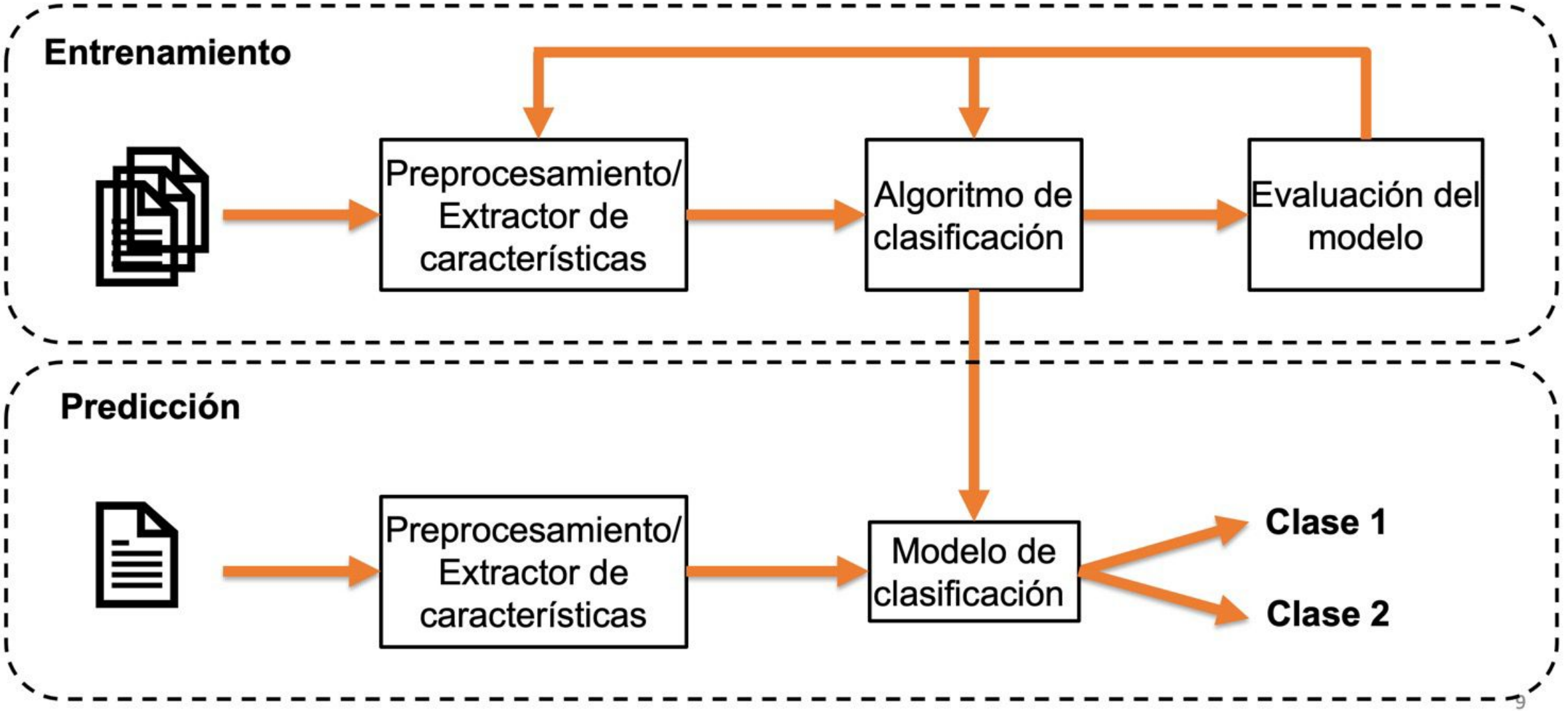
¿Qué es la ingeniería de características?

- La ingeniería de características es el proceso de transformar datos sin procesar en características informativas que son adecuadas para algoritmos de aprendizaje automático.
- Por lo general, implica crear, seleccionar o modificar características del conjunto de datos original para mejorar el rendimiento de un modelo de aprendizaje automático.
- La ingeniería de características tiene como objetivo extraer información relevante de los datos y representarla de una manera que mejore la capacidad del modelo para aprender patrones y hacer predicciones precisas.
- Su objetivo es capturar las características del texto en un vector numérico que pueda ser entendido por los algoritmos de ML. (Cf. *construcción*, *definiciones operativas* y *medición* en ciencia experimental)
- En resumen, se trata de cómo representar de manera significativa los textos cuantitativamente, es decir, la representación del texto.

Ingeniería de características para ML clásico

- Listas de frecuencias basadas en palabras
- Representaciones de bolsa de palabras
- Listas de frecuencia de palabras específicas del dominio
- Características hechas a mano basadas en el conocimiento específico del dominio

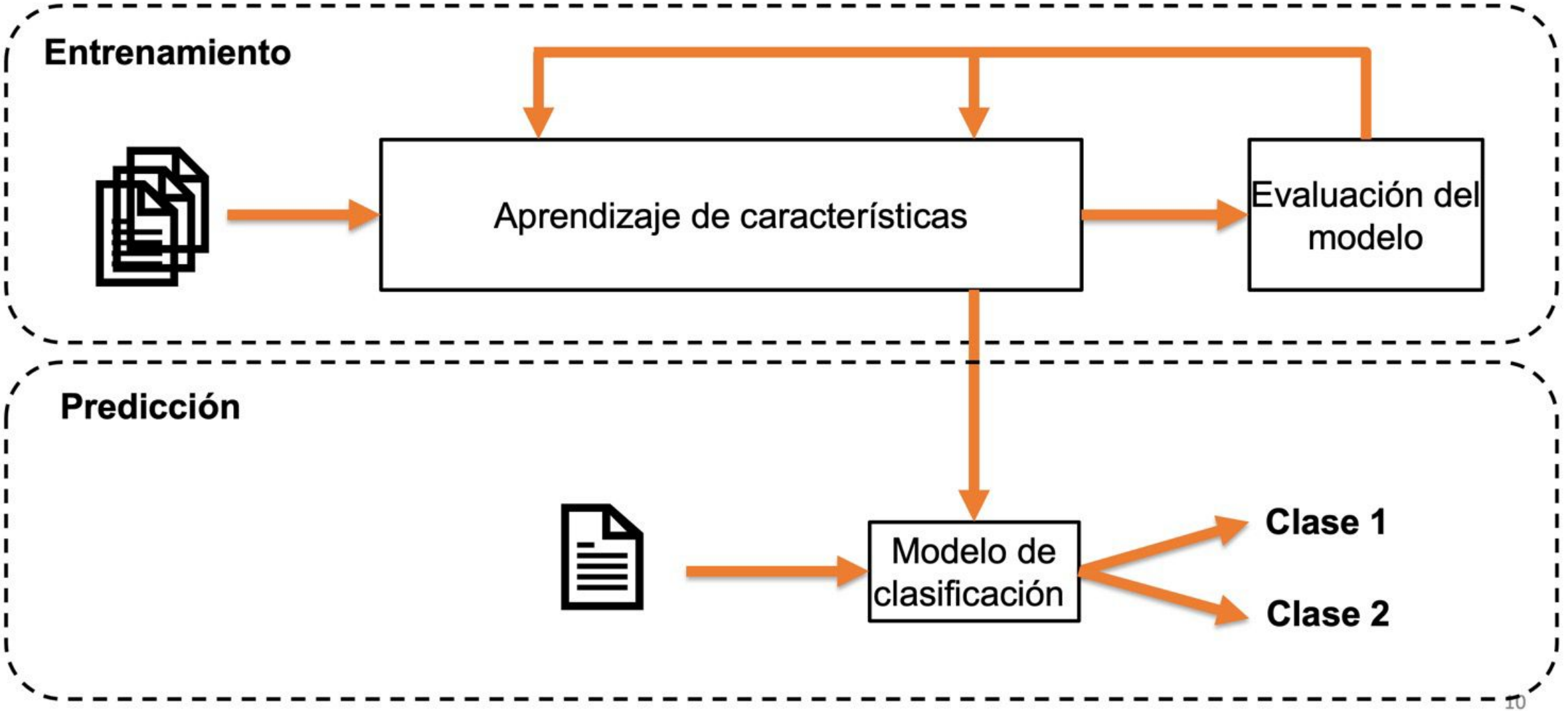
Aprendizaje de características



Ingeniería de características para DL

- DL toma directamente los textos como entradas al modelo.
- Los modelos de aprendizaje profundo, particularmente las redes neuronales profundas, son capaces de extraer automáticamente características significativas de los datos sin procesar a través de múltiples capas de abstracción.
- Este proceso se conoce como aprendizaje de características o aprendizaje de representación, y es una ventaja clave del aprendizaje profundo sobre los enfoques tradicionales de aprendizaje automático.
- Sin embargo, un inconveniente es que los modelos de aprendizaje profundo a menudo son menos interpretables en comparación con los enfoques tradicionales de aprendizaje automático.

Aprendizaje de características



Referencias

- Capítulo 1 y 2 de Practical Natural Language Processing. [Vajjala et al., 2020]