

OpenCourseWare
**Procesamiento de Lenguaje Natural con
Aprendizaje Profundo,**
Máster en Ciencia y Tecnología Informática

**Tema 7.2 - Desarrollo de aplicaciones PLN basadas
en aprendizaje profundo. Reconocimiento de
entidades (Named Entity Recognition)**

Índice


- **Recuperación de información vs Extracción de información.**
- ¿Qué es NER?, por qué es importante?
- Enfoques NER
- Retos en NER

Recuperación de Información (Information Retrieval)

Query



Documentos relevantes

 [juntaelectoralcentral.es](http://www.juntaelectoralcentral.es)
<http://www.juntaelectoralcentral.es> › Locales_mayo2023 ›

Elecciones. Elecciones en curso. 28 de mayo de 2023.

Elección en curso. **Elecciones Municipales.** 28 de mayo de **2023.** Información general · Administración electoral · Electores · Candidaturas · Mesas electorales ...

 [juntaelectoralcentral.es](http://www.juntaelectoralcentral.es)
<http://www.juntaelectoralcentral.es> › Locales_mayo2023 ›

28 de mayo de 2023 - Junta Electoral Central

Elección en curso. **Elecciones Municipales.** 28 de mayo de **2023.** Información general · Administración electoral · Electores · Candidaturas · Mesas electorales ...

 [ine.es](https://www.ine.es)
<https://www.ine.es> › dyngs › CEL ›

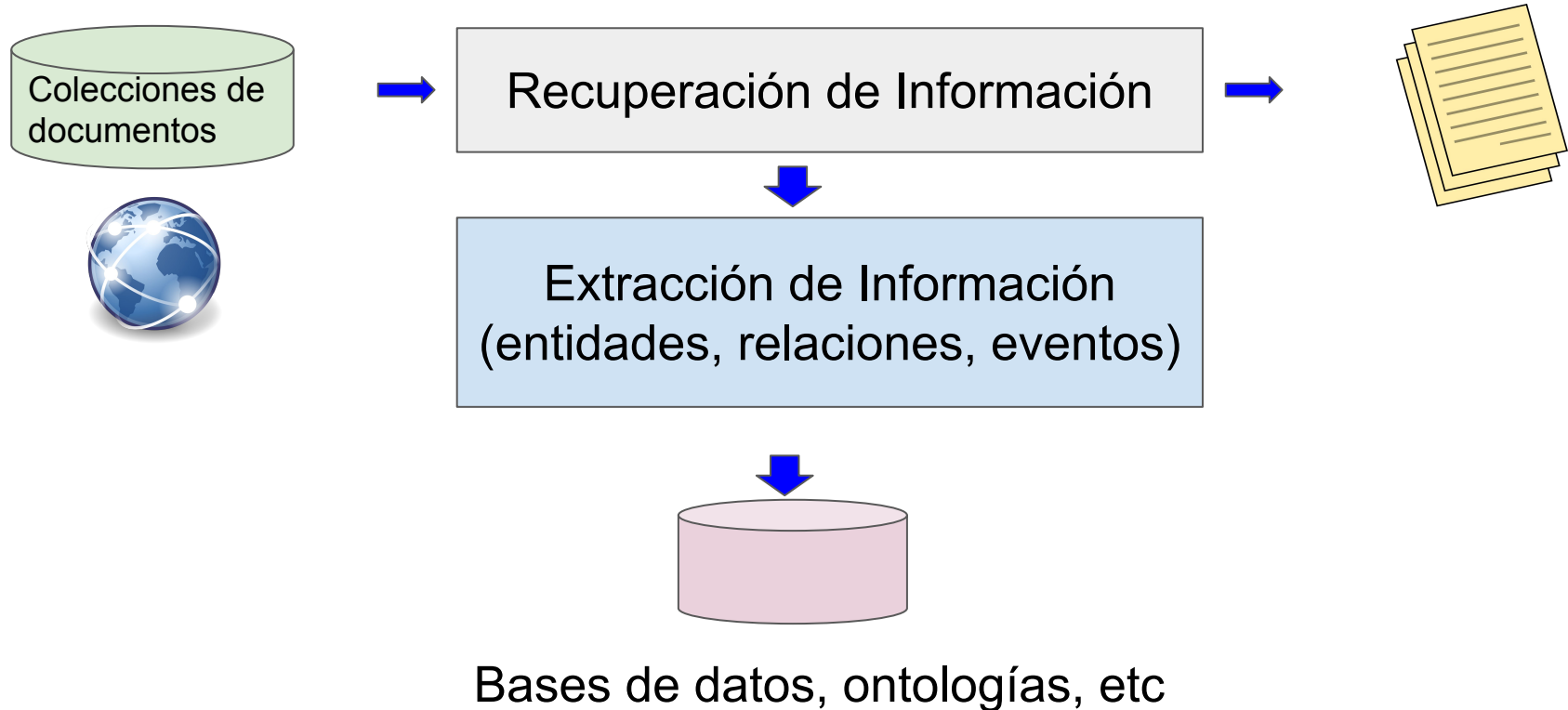
Elecciones municipales y autonómicas de 28 de mayo de 2023

Las tareas que va a desarrollar la Oficina del Censo Electoral se derivan de las competencias que en los procesos electorales le asigna la Ley Orgánica 5/1985, ...

[Censo Electoral / Elecciones...](#) · [Lista de tablas](#) · [Trámites telemáticos a...](#)

Extracción de Información

Información no estructurado (textos) -> Estructurado (bases de datos, ontologías, etc)



Extracción de Información. Tareas:

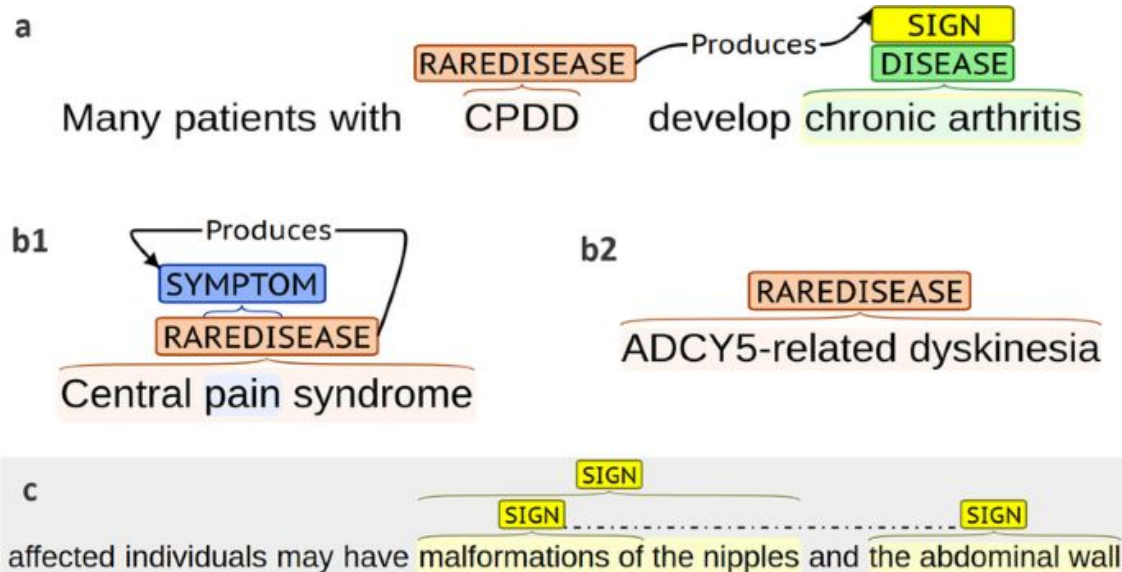
Reconocimiento de Entidades Nombradas. (Named Entity Recognition (NER))

Isabel vive en Madrid . Ella trabaja en la UC3M

PER LOC ORG

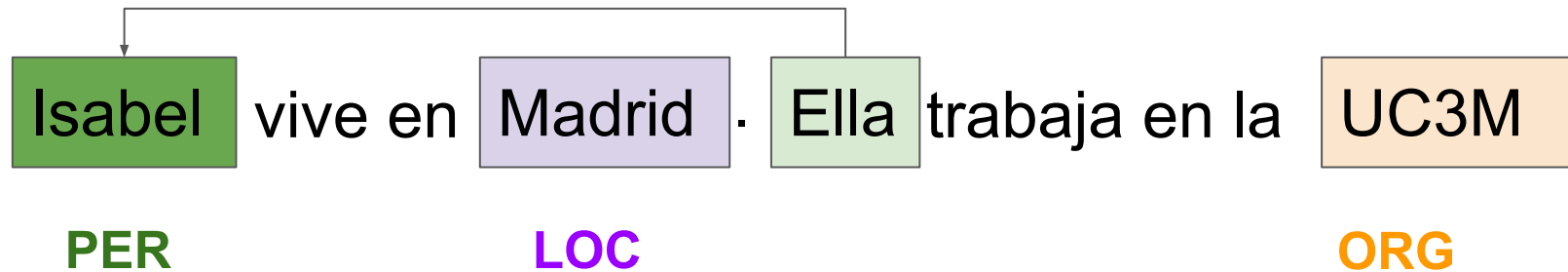
Reconocimiento de Entidades

- El conjunto de tipos de entidades depende del dominio de aplicación.



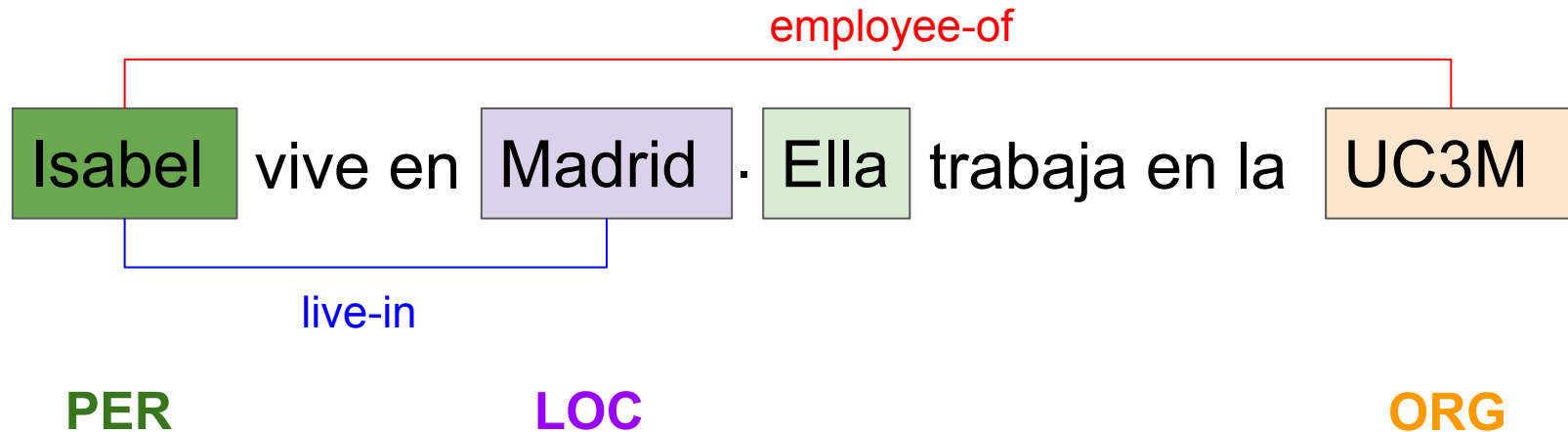
Extracción de Información. Tareas:

Resolución de la Correferencia



Extracción de Información. Tareas:

Extracción de relaciones



Índice

- Recuperación de información vs Extracción de información.
- **¿Qué es NER?, por qué es importante?**
- Enfoques NER
- Retos en NER

¿Por qué NER?

- NER es una tarea imprescindible para muchas otras tareas de PLN: recuperación de información, búsqueda de respuestas, traducción automática, etc

***¿Quién** es el actual **director** del **FMI**?*

↓
Persona

↓
Atributo de la
persona

↓
Acrónimo:
Organismo
Fondo Monetario
Internacional

Índice

- Recuperación de información vs Extracción de información.
- ¿Qué es NER?, por qué es importante?
- **Enfoques NER**
- Retos en NER

Principales enfoques para NER

- Reglas (expresiones regulares) y diccionarios (listas de nombres, organizaciones, etc)
- Aprendizaje automático: algoritmos clásicos (CRF) vs redes profundas.

Enfoques: reglas y diccionarios para NER

- Ejemplos:
 - Para identificar nombres de universidades y clasificarlas como ORG:
 - *“Universidad de “ + LOC*
 - Para identificar nombres de personas.
 - *[Mr.|Mrs.|Dr.|..” + Xxxx (primer carácter es una letra mayúscula y el resto son letras minúsculas).*
- Alta precisión, bajo recall.

Enfoques: aprendizaje automático para NER

- La tarea se plantea como una tarea de clasificación de tokens (sequence labeling):
 - El texto es tokenizado y cada token es clasificado con una etiqueta (standard IOB).
- El objetivo es entrenar un modelo capaz de clasificar cada token en un oración.

Enfoques: aprendizaje automático para NER

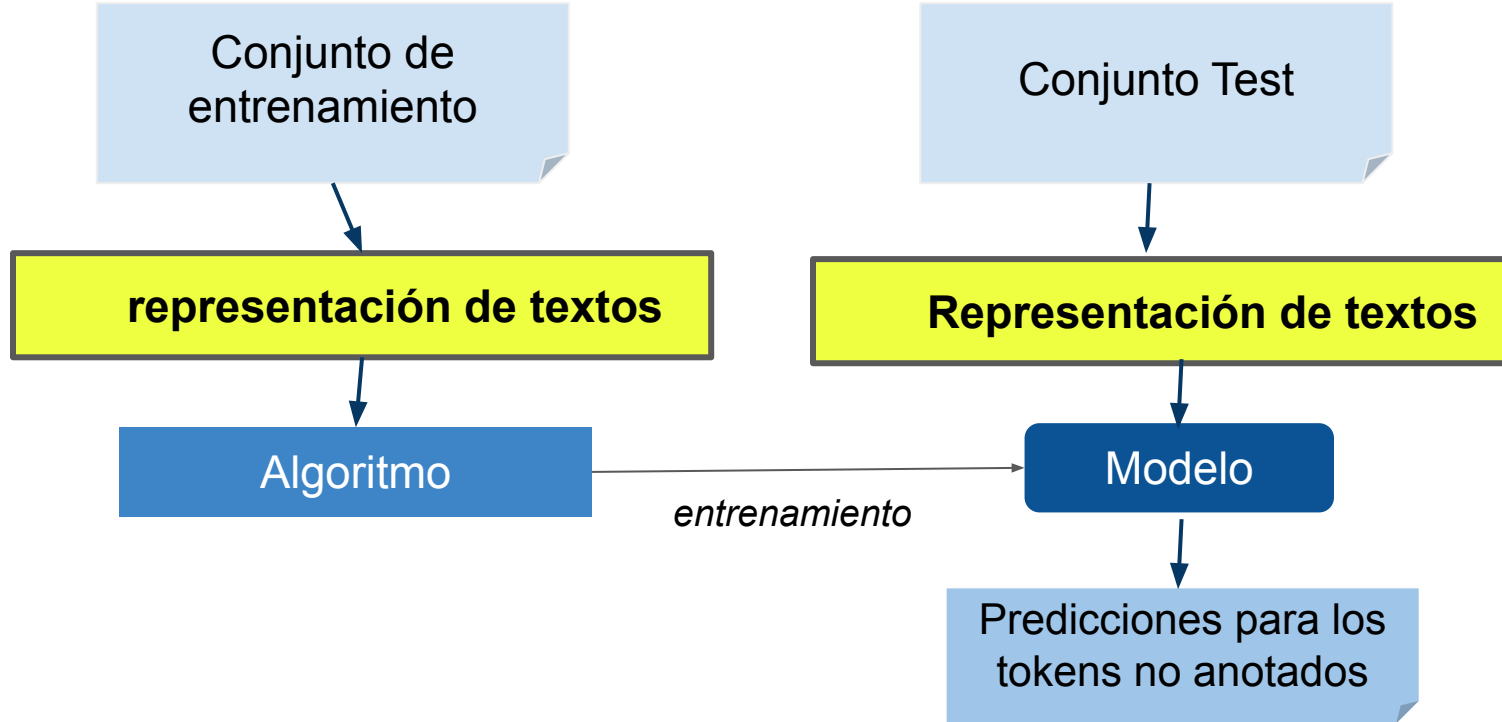
- Standard IOB
(inside-outside-beginning):
Conjunto de etiquetas: **O**,
B-X, **I-X**, donde X es un tipo
de entidades (LOC, ORG,
PER, etc).
 - **O** indica que es un token
que no pertenece a una
entidad.
 - **B** indica que es el primer
token de una entidad
 - **I** indica que es un token
interno de una entidad

Token	Etiqueta	Tipo de Entidad
Yolanda	B	PER
Díaz	I	PER
visitó	O	-
hoy	O	-
la	O	-
Comisión	B	ORG
Europea	I	ORG
.	O	-

Enfoques: aprendizaje automático para NER

- basado en **características** (algoritmos clásicos como CRF).
- basado en **aprendizaje profundo** (no es necesario definir el conjunto de características).

Enfoques: aprendizaje automático para NER



Enfoques: aprendizaje automático para NER

- Feature engineering para NER:
 - Es necesario definir un conjunto de características para representar cada instancia (token).
 - El texto es preprocesado para representar cada instancia (token) con el conjunto de características definido.
 - Se entrena un algoritmo clásico (CRF, SVM, etc)

Enfoques: aprendizaje automático para NER

- Las características (features más habituales) para el problema de NER son:
 - Token y lema o stem:
 - *Isabel (token), Isabel (lema)*
 - Categoría Morfosintáctica (PoS tag):
 - *NNP (nombre propio)*
 - Patrón ortográfico de una palabra:
 - *Isabel -> Xxxxxxx*
 - Afijos (prefijos y sufijos): dependiendo del dominio pueden existir un conjunto de afijos que nos permitan identificar algunas entidades.
 - Por ejemplo, *los nombres de antidepresivos suelen terminar con el sufijo “-oxetina” (paroxetina, fluoxetina)*
 - Si el token está presente o no en un diccionario o base de datos relacionada con el tipo de entidad a identificar. Es un valor booleano.

Enfoques: aprendizaje automático para NER

- Información de contexto:
 - “La profesora **Isabel** Segura enseña *PLN*”,

Enfoques: aprendizaje automático para NER

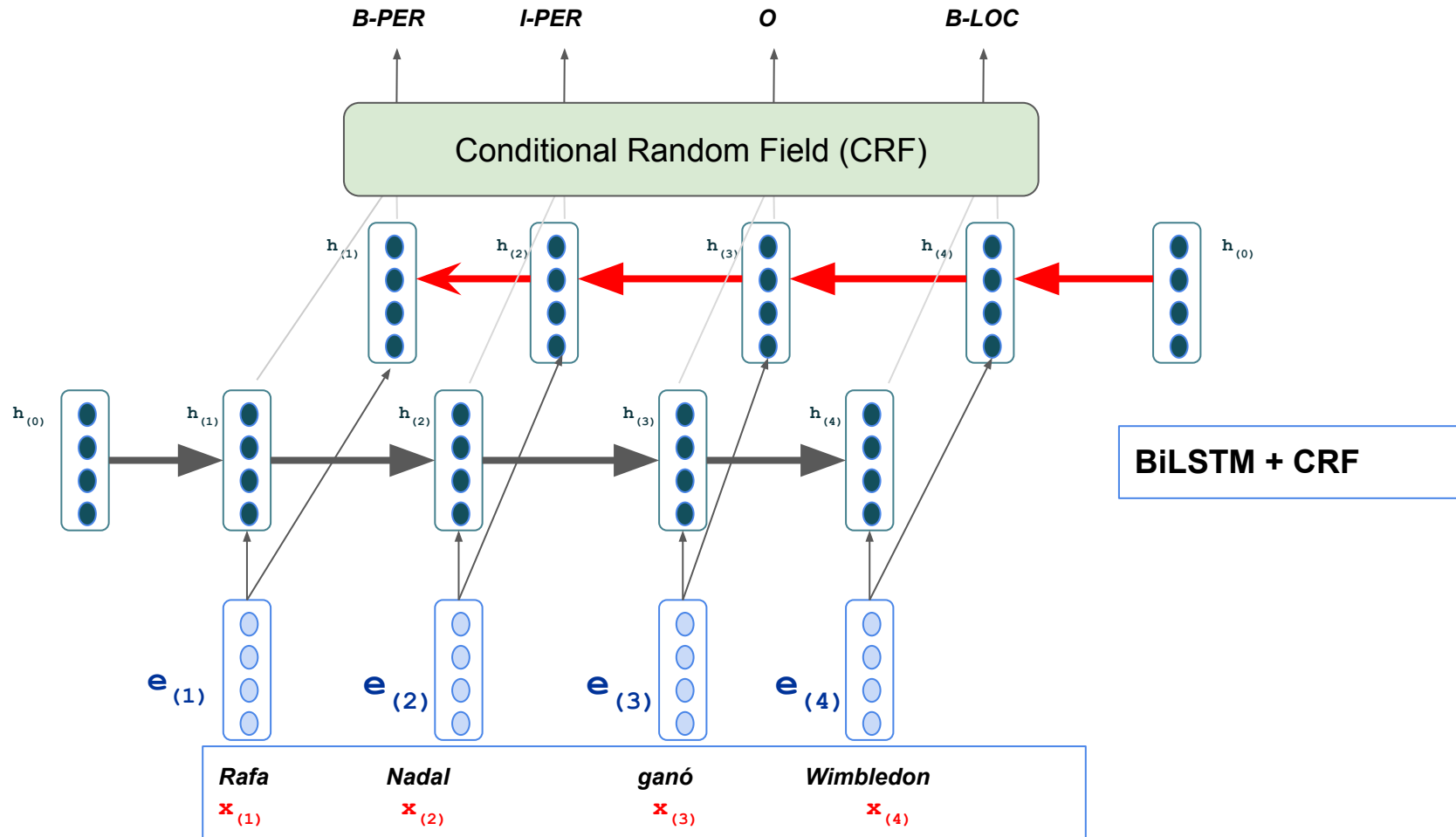
- **Información de contexto:**

- También es muy importante la información de los tokens que rodean al token a clasificar.
- Se suele considerar ventanas de tamaño 2 o 3. Por ejemplo, si estamos tratando de representar el token “Isabel” en la oración “La profesora Isabel Segura enseña NLP”, y estamos considerando ventanas de tamaño 2, también deberíamos representar la información de los tokens: La, profesora, Segura, enseña.
- De cada token en dicha ventana, se suelen obtener las características anteriormente citadas (token, lema, afijo, patrón ortográfico, etc)

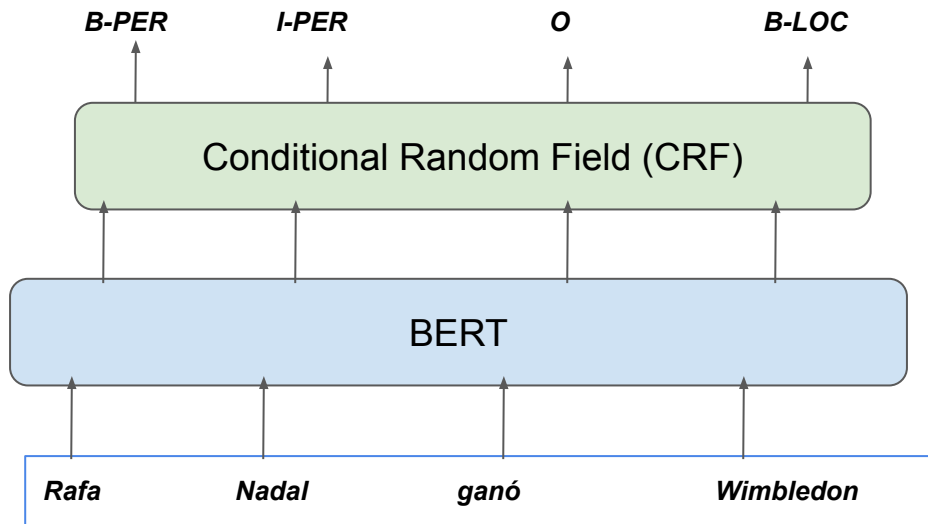
Enfoques: aprendizaje automático para NER

- En los modelos de aprendizaje profundo, el texto únicamente debe ser tokenizado. No es necesario representar cada token con un conjunto de características.
- La red profunda será capaz de aprender durante el entrenamiento las características (vectores) más apropiados para cada token en la tarea.
- Tanto en los enfoques basados en algoritmos clásicos (CRF) como en los enfoques basados en aprendizaje profundo, siempre vamos a necesitar un dataset de entrenamiento y test donde cada token esté anotado con su etiqueta IOB ([slide 15](#))

Enfoques: aprendizaje automático para NER



Enfoques: aprendizaje automático para NER



fine-tuning sobre el dataset para NER

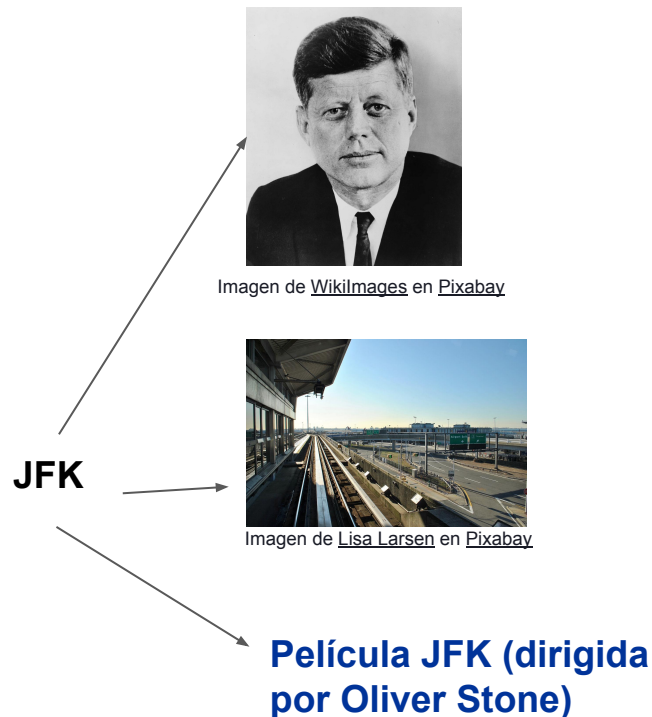
modelo pre-entrenado. El tokenizador de BERT tokeniza los textos y prepara el formato que necesita BERT

Índice

- Recuperación de información vs Extracción de información.
- ¿Qué es NER?, por qué es importante?
- Enfoques NER
- **Retos en NER**

Retos en NER

- Una misma **entidad** puede aparecer con **distintas menciones**:
 - *Kennedy, JFK, J. Kennedy, President Kennedy.*
- Una **misma mención** puede ser clasificada con **distintos tipos de entidad**
 - *JFK puede ser clasificado como PERSON, PLACE, Movie, etc...*
- Una **mención** puede referirse a dos **entidades distintas**:
 - *“Almeida visita la Sexta Noche”... ¿es Cristina Almeida, o el alcalde de Madrid, José Luis Martínez-Almeida?*



Retos en NER

- Anidamiento de entidades:
 - **Diabetes Mellitus tipo 1.**
 - **Diabetes Mellitus tipo 2.**
 - **Diabetes Gestacional.**
 - **Diabetes tipo MODY.**
 - **Diabetes tipo LADA.**
- Entidades discontinuas:
 - *Los tipos de **cáncer** más comunes son los **de mama, pulmón, colon y próstata.***

Retos en NER

- Las entidades también podrían estar representados por frases:

Affected individuals develop characteristic SIGN loss of body fat (adipose tissue)

OpenCourseWare
Procesamiento de Lenguaje Natural con
Aprendizaje Profundo,

Gracias!!!

<https://github.com/iseaura>