

# 4 - Extracción de Información: Reconocimiento de entidades y Extracción de Relaciones

# Motivación



- Exponencial crecimiento y disponibilidad de datos.
  - 2013, 3.5 ZB
  - 2022, 40 ZB
  - 2025, 180 ZB
- **Más de un 80%, datos no estructurados y en formato texto**

# 1 ZB = 1 trillón de GB

# ¿Qué es Extracción de Información?

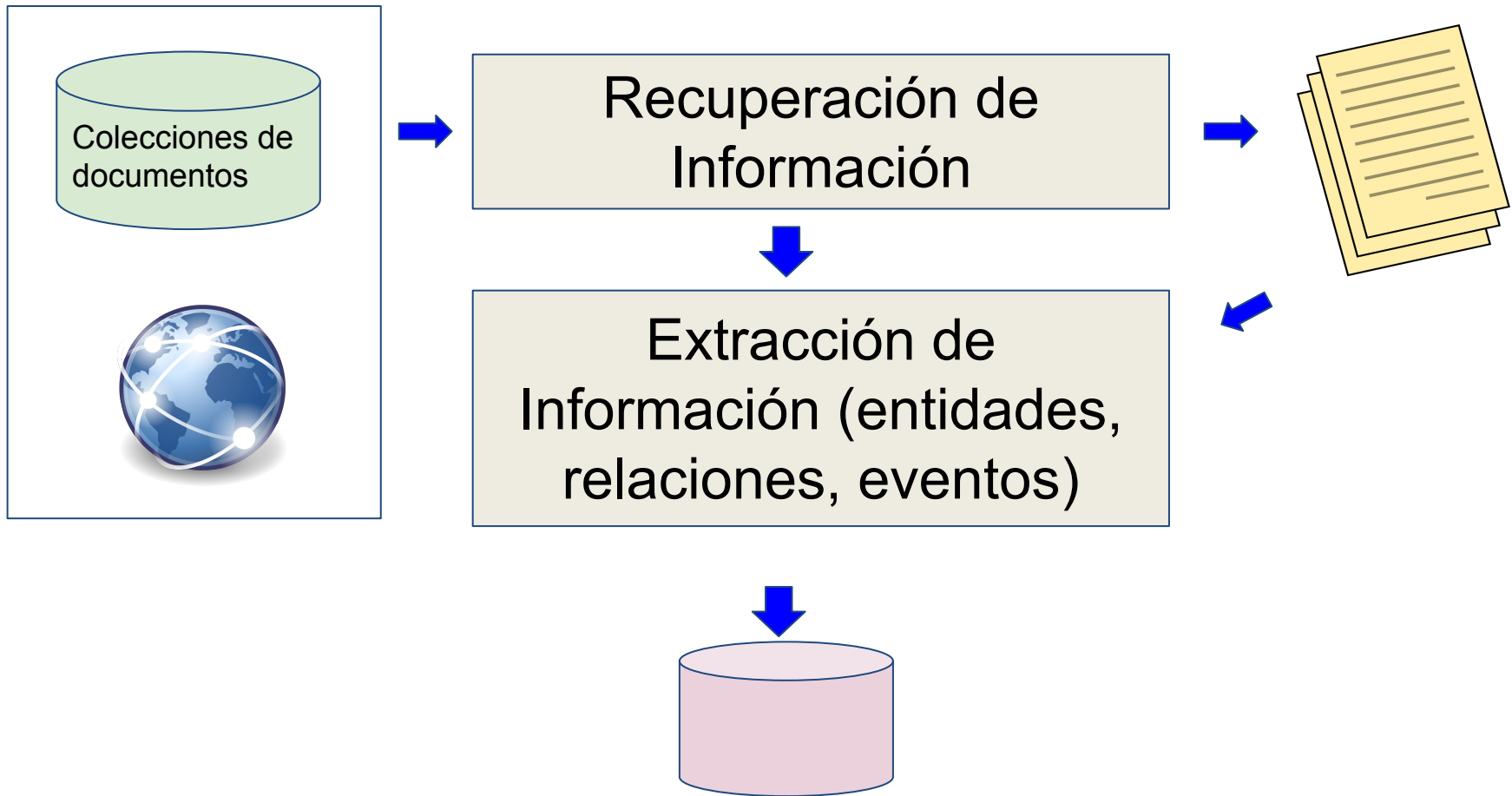
- Tarea del Procesamiento del Lenguaje Natural (PLN) que consiste en obtener información estructurada a partir de un textos no estructurados.

# ¿Por que IE?

- ¿Qué es el asma? ¿Cuáles son los principales síntomas del asma?
- ¿Qué es 2019-nCoV? ¿Origen del coronavirus?, ¿países afectados?, ¿número de fallecidos?, ¿número de contagios por país?, ¿medidas para evitar el contagio, etc.
- ¿Donde nació Rafa Nadal? ¿cuántos torneos ha ganado? ¿cuál ha sido su última victoria?, ¿cuál ha sido su última victoria?
- ¿Número de desempleados en España?, ¿provincias con mayor desempleo?, ...

# Fases IE

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



Bases de datos, ontologías, etc

# Tareas - Extracción de Información

1. Reconocimiento de entidades (Named Entity Recognition - NER).
2. Resolución de correferencias.
3. Extracción de relaciones

# Reconocimiento de entidades

**¿Qué son entidades?** - Nombres propios.

- Personas (PER): *Manuela Carmena, James Bond...*
- Organizaciones (ORG): *Organización de las Naciones Unidas, IBM SA, PSOE, AMPA...*
- Lugares (LOC): *Madrid, Puerta de Toledo, España, Avenida de la Universidad...*
- Fechas (*8 de Marzo*),
- ...

# Reconocimiento de entidades

- Los tipos de entidades dependen del dominio:
  - Productos, marcas: *Coca-Cola, iPhone, Nocilla.*
  - Enfermedades y síntomas: *diabetes, cancer de colón, ictus, asma, eritema, hepatitis C, etc.*
  - Sustancias químicas: *cloruro de sodio, NaCL, etanol, etc*
  - Fármacos: *paroxetina, lorazepam, penicilina, etc.*
  - Genes y proteínas: *TP53, TNF, EGFR, VEGFA, APOE, IL6, TGFB1, MTHFR, ESR1 y AKT1*
  -



# Reconocimiento de entidades

1. Paso previo para la tarea de Extracción de relaciones (*genes y enfermedades, fármacos y efectos adversos, entrenadores y equipos de futbol*, etc)
2. Tarea imprescindible para un gran número de aplicaciones PLN (búsqueda de información de textos, análisis de sentimiento, anonimización de textos, traducción automática, etc).

# ¿Por qué NER?

*Quién es el actual director del FMI?*

# ¿Por qué NER?

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



Persona



Atributo de la  
persona



Acrónimo:  
Organismo  
Fondo Monetario  
Internacional

# ¿Por qué NER?

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



- Fondo Monetario Internacional
- Finland Meteorological Institute
- Fiasco Mundial de Inútiles

# Anonimización de Notas Clínicas

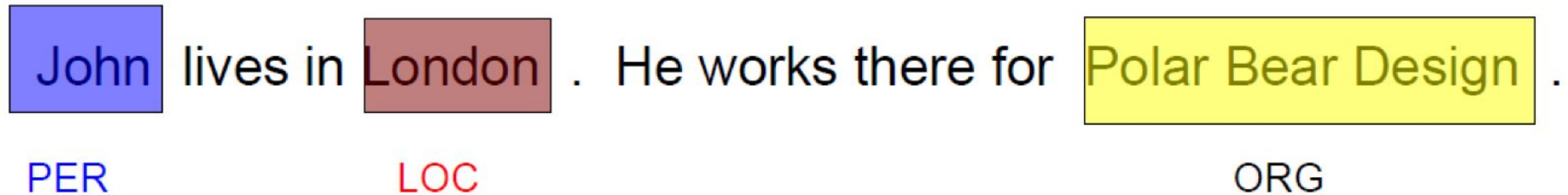
1	Datos del paciente.	
2	Nombre: <span>NOMBRE SUJETO ASISTENCIA</span> Pedro.	
3	Apellidos: <span>NOMBRE SUJETO ASISTENCIA</span> Jimenez Ramos.	
4	NHC: <span>N SUJETO ASISTENCIA</span> 4763954.	
5	NASS: <span>ID ASEGURAMIENTO</span> 47 37584930 84.	
6	Domicilio: <span>CALLE</span> Calle del pez, 28.	
7	Localidad/ Provincia: <span>TERRITORIO</span> Madrid.	
8	CP: <span>TERRITORIO</span> 28001.	
9	Datos asistenciales.	
10	Fecha de nacimiento: <span>FECHAS</span> 20/05/2000.	
11	País: <span>PAIS</span> España.	
12	Edad: <span>EDAD SUJETO ASISTENCIA</span> 16 años Sexo: <span>SEXO SUJETO ASISTENCIA</span> H.	
13	Fecha de Ingreso: <span>FECHAS</span> 26/08/2017.	
14	Servicio: Urgencias.	
15	Médico: <span>NOMBRE PERSONAL SANITARIO</span> Luis Moyano Calvo N°Col: <span>ID_TITULACION_PERSONAL_SANITARIO</span> 28 31 23567.	
16	Informe clínico del paciente: <span>EDAD SUJETO ASISTENCIA</span> Adolescente <span>SEXO SUJETO ASISTENCIA</span> Varón de <span>EDAD SUJETO ASISTENCIA</span> diecisiete años sin antecedentes de interés que acude por	
17	En la analítica de orina se aprecian 30-50 hematíes por campo. Urocultivo negativo.	
18	Se practica ecografía abdominal observándose pequeña lesión de medio centímetro de diámetro, sólida con refuerzo hiperecogénico anterior.	
19	Realizamos cistoscopia observándose en cara lateral derecha, por fuera de orificio ureteral dos pequeñas lesiones sobre elevadas, con muco:	
20	Sospechándose lesión inflamatoria se prescribe tratamiento con A.I.N.E. durante diez días sin que desaparezcan las lesiones, decidiéndose in	
21	Se realiza RTU de ambas lesiones vesicales, siendo el informe anatomopatológico el de leiomioma vesical, describiendo la lesión como "pro eosinófilo sin atipia, necrosis ni actividad mitótica significativa. Con el estudio inmunohistoquímico se demostró intensa positividad citoplasmá	
22	Remitido por: Dr. <span>NOMBRE PERSONAL SANITARIO</span> Luis Moyano Calvo <span>CALLE</span> C/ Eduardo Rivas, 3 <span>TERRITORIO</span> 28018 <span>TERRITORIO</span> Madrid. <span>PAIS</span> España. e-mail: <span>CORREO ELECTRONICO</span> joseluis Moyano@ya.com	

# Reconocimiento de Entidades

John lives in London . He works there for Polar Bear Design .

# Reconocimiento de Entidades

John lives in London . He works there for Polar Bear Design .



# Reconocimiento de Entidades

ORG

Destacados representantes del Parlamento y la prensa rusos criticaron hoy el "belicismo" que ha definido como posible blanco de su lucha antiterrorista.

ORG

PER

El presidente de la Duma (cámara baja), Guennadi Selezniyov, calificó de "claramente apor-

ORG

LOC

PER

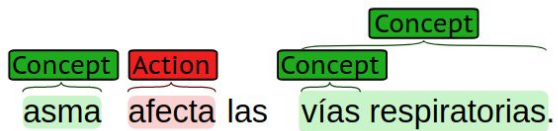
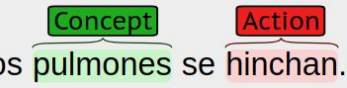
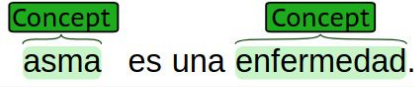
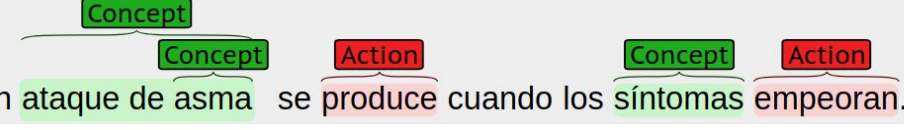

del Kremlin para Chechenia, Serguéi Yastrzhembski.

LOC

El asesor presidencial dijo que Rusia puede lanzar un ataque preventivo contra los camp



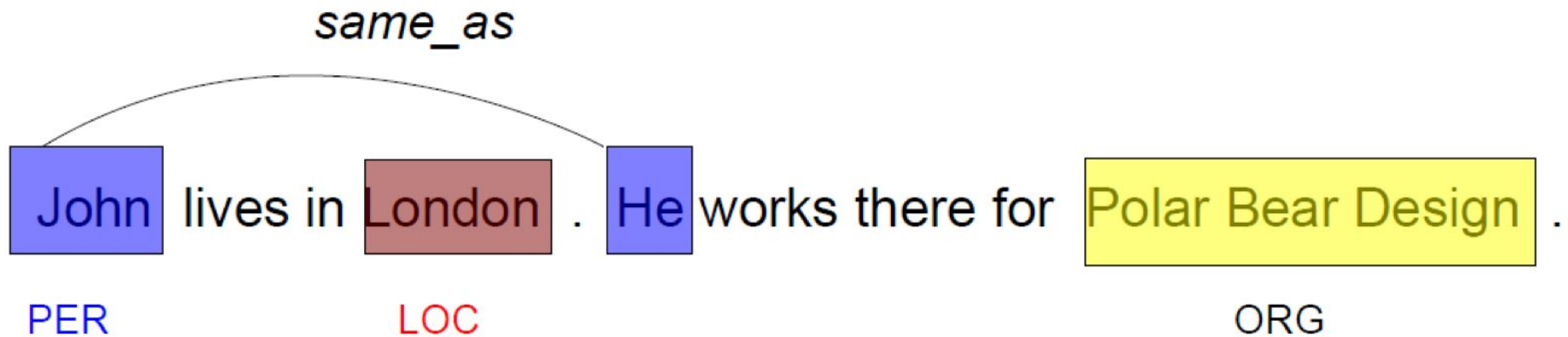
# Reconocimiento de Entidades

- 1 El asma afecta las vías respiratorias.  

- 2 Los pulmones se hinchan.  

- 3 El asma es una enfermedad.  

- 4 Un ataque de asma se produce cuando los síntomas empeoran.  

- 5 Los síntomas y el tratamiento dependen del tipo de cáncer y de lo avanzada que esté la enfermedad.  


# Tareas - Extracción de Información

1. Reconocimiento de entidades (Named Entity Recognition - NER).
2. **Resolución de correferencias.**
3. Extracción de relaciones

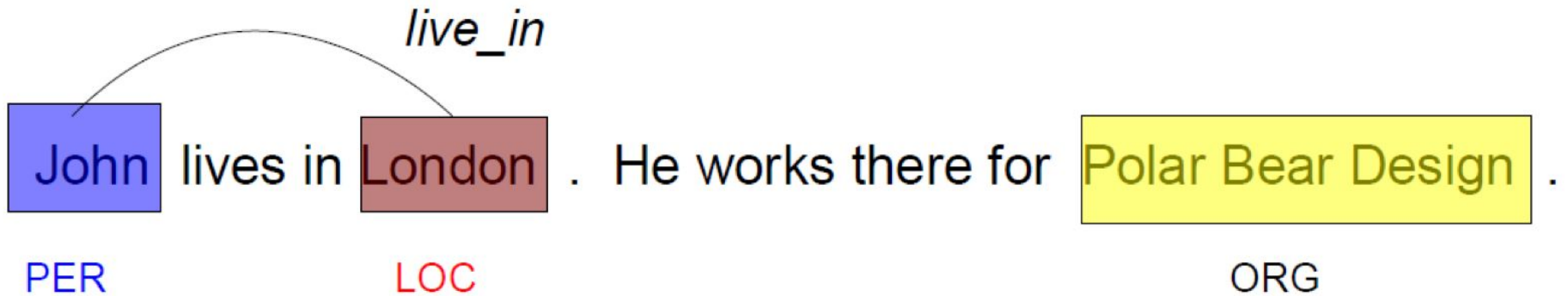
# Resolución de correferencias



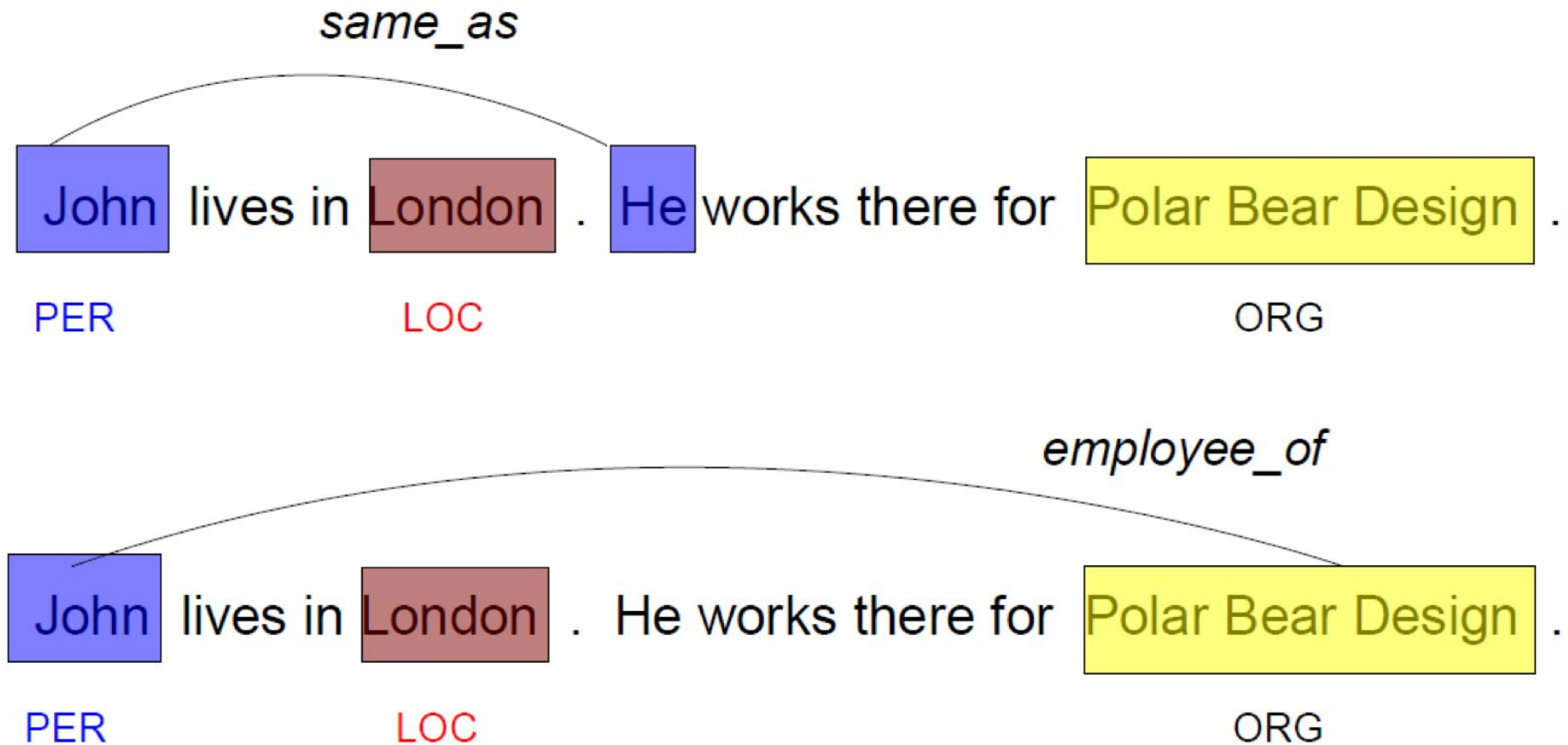
# Tareas - Extracción de Información

1. Reconocimiento de entidades (Named Entity Recognition - NER).
2. Resolución de correferencias.
3. **Extracción de relaciones**

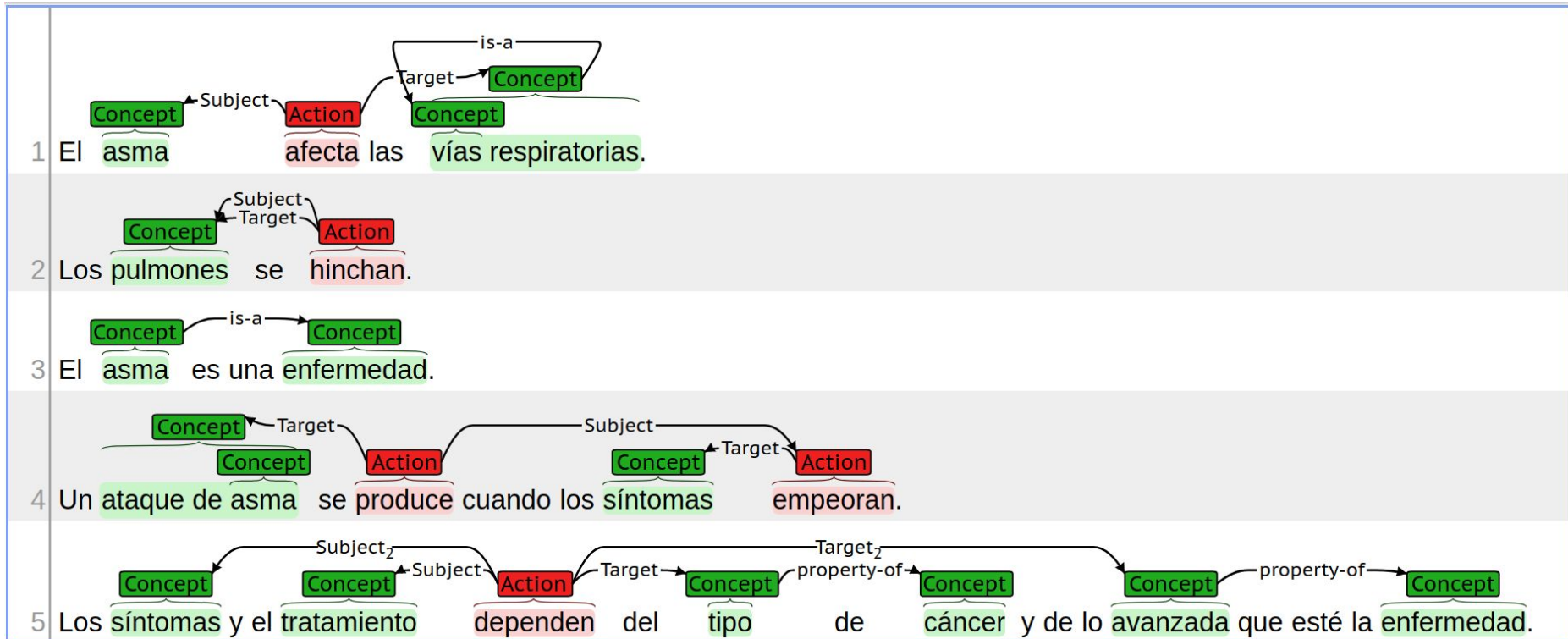
# Extracción de Relaciones



# Extracción de Relaciones



# Extracción de Relaciones



# Más preguntas

- ¿Quiénes han sido los directores del FMI durante los 20 últimos años?
- ¿Lista de los posibles efectos adversos de la paroxetina?
- ¿Dónde está la sede de la empresa Inditex?
- ¿Cuándo se fundó el Banco Central Europeo?
- ¿Edad media de la aparición del autismo?



# Extracción de Información

**Input:** Apple took its annual spring event to Chicago this year.

## Tokenization

Apple / took / its / annual / spring / event / to / Chicago / this / year

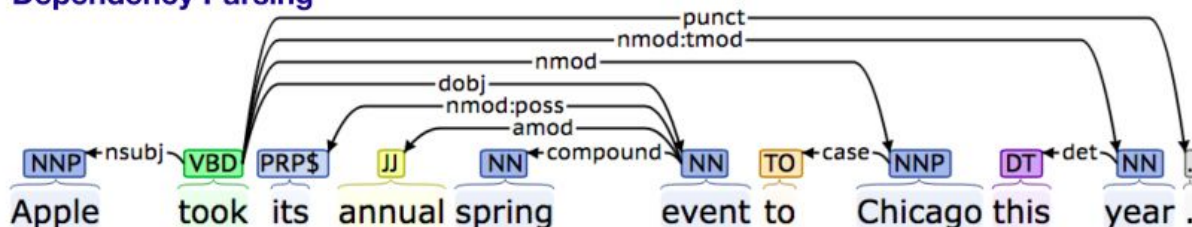
## Part-of-Speech Tagging

NNP VBD PRP\$ JJ NN NN TO NNP DT NN .  
Apple took its annual spring event to Chicago this year .

## Lemmatization

Apple take its annual spring event to Chicago this year .  
Apple took its annual spring event to Chicago this year .

## Dependency Parsing



<https://iccl.inf.tu-dresden.de/w/images/b/b9/Lecture-03.pdf>

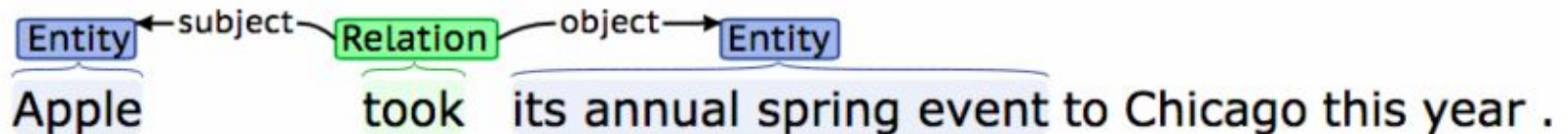
# Extracción de Información

**Input:** Apple took its annual spring event to Chicago this year.

## Named Entity Recognition



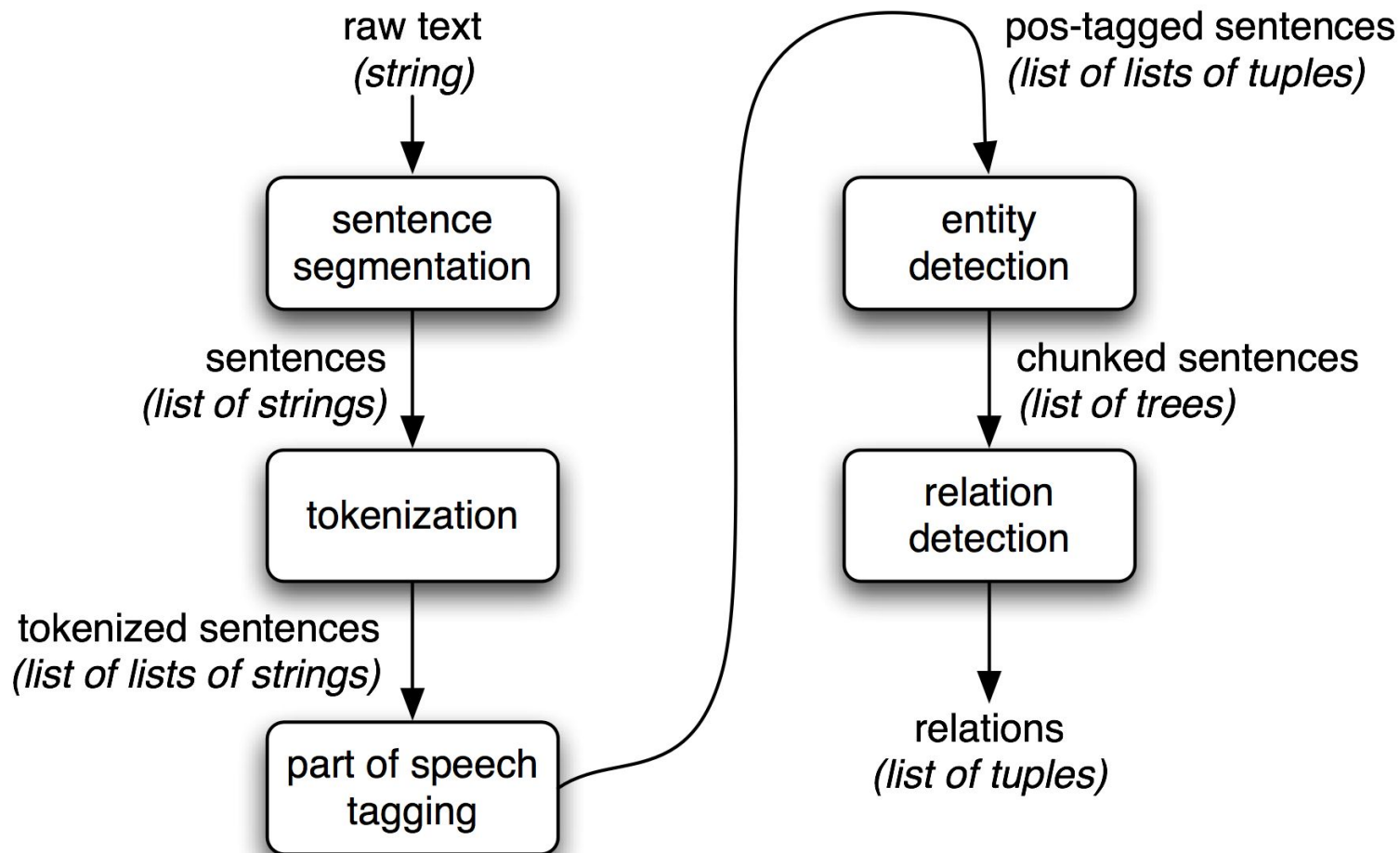
## Relation Extraction



## Coreference Resolution



# Arquitectura de un sistema Extracción de Información



# Evaluación

- Basada en Precisión, Recall y F1.

# ¿Cómo evaluar?

		Actual	
		Positive	Negative
Predicted	Positive	<b>True Positive</b>	<b>False Positive</b>
	Negative	<b>False Negative</b>	<b>True Negative</b>

<https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>

# ¿Cómo evaluar? - Precision

		Actual	
		Positive	Negative
Predicted	Positive	<b>True Positive</b>	<b>False Positive</b>
	Negative	<b>False Negative</b>	<b>True Negative</b>

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

<https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>

# ¿Cómo evaluar? - Recall

		Actual	
		Positive	Negative
Predicted	Positive	<b>True Positive</b>	<b>False Positive</b>
	Negative	<b>False Negative</b>	<b>True Negative</b>

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FN})$$

<https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>

# ¿Cómo evaluar? - F1

		Actual	
		Positive	Negative
Predicted	Positive	<b>True Positive</b>	<b>False Positive</b>
	Negative	<b>False Negative</b>	<b>True Negative</b>

$$\mathbf{F1 = 2 * Precision * Recall / (Precision + Recall)}$$

<https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>