

Extraccion de Informacion

☐ https://drive.google.com/file/d/1XTHJSRSdyV3YxKKxybJI-cG0NKeeFg_R

☐ <https://drive.google.com/file/d/15mRHicaSndmRx28Ezc1uh2U6rzXjaPqI/view>

Métodos supervisados y auto-supervisados

Open Information Extraction - Extracción de información

Resumen

Reconocimiento de nombres de entidades (NER)

Modelos de etiquetamiento secuencial para el reconocimiento de nombres de entidades (NER)

NER: Pasos del entrenamiento

Etiquetamiento secuencial

IO - encoding

IOB - encoding

NER: Identificación de características

Algoritmos de inferencia

El objetivo de la extracción de información es capturar ciertas **partes relevantes** de un **texto**.

Muchas veces en el contexto de varios documentos distintos

Generar con dicha información una representación estructurada, limpia y legible, como podría ser una tupla en una base de datos relacional.



Información factica → info sobre hechos

¿Quién hizo qué, a quién y cuándo?

Ejemplo: Las oficinas de Google en la Argentina ya tienen su historia. La empresa abrió su filial local en 2008 en Puerto Madero. Allí trabajan 215 empleados en los 6000 m2 que ocupan las instalaciones.

Ejemplo:

Las oficinas de Google en la Argentina ya tienen su historia. La empresa abrió su filial local en 2008 en Puerto Madero. Allí trabajan 215 empleados en los 6000 m2 que ocupan las instalaciones.

- SEDE("Google_Arentina", "Puerto Madero")
- APERTURA_SEDE("Google_Arentina", "2008")

Me falta la relacion entre esas entidades, sustantivos → abrio y filial local

Correr un algoritmo que detecte los dos valores sede y apertura_sede

Historia

1991: DARPA

"Construir sistemas robustos capaces de llenar plantillas con piezas de conocimiento sobre el terrorismo en América Latina"

- Fechas
- Ubicaciones
- Perpetradores
- armas
- víctimas
- objetivos físicos

Lo que se hizo fue construir patrones, expresiones regulares o reglas de patrones de coincidencia

Para fines de los 90' los dominios habían cambiado:

- joint ventures
- microelectrónica
- sucesión de gestiones empresariales



Este tipo de sistemas estaban basados en reglas de coincidencia de patrones creados a mano

Métodos supervisados y auto-supervisados

Supervisados:

- Requieren un conjunto de datos previamente etiquetados
- Requieren tiempo, esfuerzo y una intervención humana importante

Auto-Supervisados:

- Aprenden a etiquetar y generar su propio conjunto de entrenamiento
- Son escalables

Open Information Extraction - Extracción de información

2007: Métodos de extracción para la Web (Open Information Extraction)

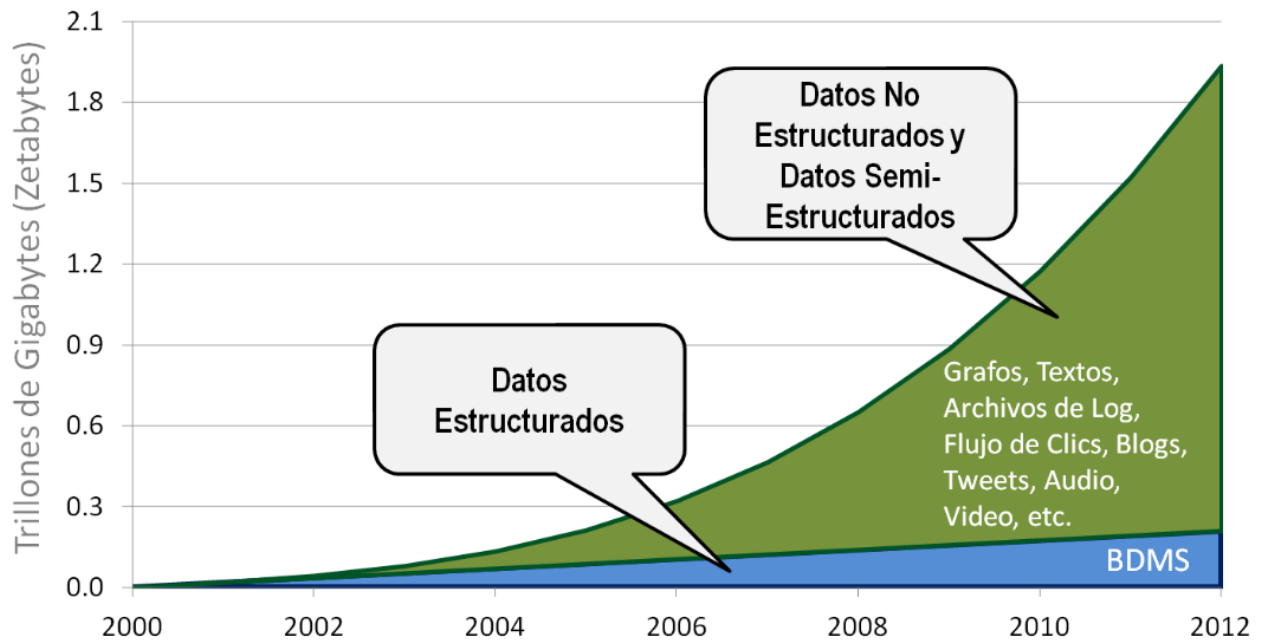
- No supervisados
- Independientes del dominio
- Trabajan con grandes cantidades de datos (corpus)

Ejemplo:



Detección automática de eventos en un e-mail de Apple

Sugiere agregarlo al calendar extrayendo fechas



Resumen

Para hacer una extracción de información primero aplicamos una técnica llamada Name Entity Recognition NER para detectar entidades

Entidades → son nombres propios

El segundo paso es extraer la relación que hay entre estas dos entidades. Podemos usar varias cosas

Reconocimiento de nombres de entidades (NER)

Ready Player One es un libro de ciencia ficción. Es la primera novela de Ernest Cline. El libro fue publicado en España en noviembre de 2011 por Ediciones B. Es el año 2044 y el mundo es un desastre. Las fuentes de energía fósiles están prácticamente agotadas y el precio del combustible está por las nubes. En medio de una enorme depresión a nivel mundial la mayoría de la gente subsiste como puede. Sin embargo un videojuego de realidad virtual llamado OASIS proporciona ...

Hay diferentes nombres propios y años

Nombres de entidades → cualquier cosa que sea susceptible de tener un nombre

Sustantivos genericos no, pero nombres propios o años (nombre de un año específico)

Ready Player One es un libro de ciencia ficción. Es la primera novela de Ernest Cline. El libro fue publicado en España en noviembre de 2011 por Ediciones B. Es el año 2044 y el mundo es un desastre. Las fuentes de energía fósiles están prácticamente agotadas y el precio del combustible está por las nubes. En medio de una enorme depresión a nivel mundial la mayoría de la gente subsiste como puede. Sin embargo un videojuego de realidad virtual llamado OASIS proporciona ...

Ready Player One => Libro

Ernest Cline. => Persona

España => Lugar, país

2011 => Fecha, año

Ediciones B. => Empresa, editorial

2044 => Fecha, año

OASIS => OTRO (Vídeo Juego de ficción)

Estas tareas de NER se usan para:

- Índices o enlaces a contenidos relacionados.
- Destinatario de los sentimientos en *Sentiment Analysis*
- Extracción de información
- Preguntas y respuestas (*question answering*)

1er paso: Tokenización

- Ready => libro
- Player => libro
- One => Libro
- Ernest => Persona
- Cline. => Persona
- España => Lugar, país
- 2011 => Fecha, año

La **tokenización** consiste en dividir un texto en entidades más pequeñas llamadas tokens. Los tokens son cosas diferentes dependiendo del tipo de tokenizador que se utilice. Un token puede ser una palabra, un carácter o una subpalabra (por ejemplo, en la palabra inglesa "higher", hay dos subpalabras: "high" y "er"). Los signos de puntuación como "!", ".", y ";", también pueden ser tokens.

Ejemplo:

"*First Bank of Chicago announced earnings...*" (Primer Banco de Chicago anunció ganancias...)

First Bank of Chicago => ?

Bank of Chicago => ?

First esta con mayuscula y eso puede generar una duda porque las palabras que empiezan con mayuscula son las candidatas a ser nombre de entidades

Como detecta donde empieza o termina la entidad?

Si el algoritmo de NER detectase la segunda opción *Bank of Chicago* en vez de la primera opción, esta cometiendo una doble falta → detecta una entidad que no existe y tengo un missing de la primera entidad

Error es **doble**: falta una entidad y se añade una inexistente

Modelos de etiquetamiento secuencial para el reconocimiento de nombres de entidades (NER)

NER: Pasos del entrenamiento

1. Conseguir un conjunto de documentos representativos de nuestro dominio.
2. Etiquetar cada palabra (token) con la clase que le corresponde (persona, organización, año, lugar, etc.) o bien marcarla con la etiqueta: "otra".
3. Especificar características de extracción que se adecuen a las clases y el texto que tenemos.
4. Entrenar un clasificador secuencial para predecir las etiquetas del conjunto de prueba.

Etiquetamiento secuencial

IO - encoding

Ejemplo: ***Fred showed Sue Mengqiu Huang's new painting***

(Fred le mostro a Sue la nueva pintura de Mengqiu Huang)

PER: persona

O: otra

| Tokens | IO - encoding |
|----------|---------------|
| Fred | PER |
| Showed | O |
| Sue | PER |
| Mengqiu | PER |
| Huang | PER |
| 's | O |
| new | O |
| painting | O |

El problema con IO-encoding es cuando hay dos nombres pegados y es difícil para el algoritmo detectar cuando termina uno y empieza el otro

IOB - encoding

| Tokens | IO - encoding | IOB - encoding |
|----------|---------------|----------------|
| Fred | PER | B-PER |
| Showed | O | O |
| Sue | PER | B-PER |
| Mengqiu | PER | B-PER |
| Huang | PER | I-PER |
| 's | O | O |
| new | O | O |
| painting | O | O |

Usa dos etiquetas por clase

B: begin

I: continua la misma entidad

NER: Identificación de características

Basadas en las palabras

- palabra actual
- palabra previa o siguiente
- substring de una palabra
- forma de una palabra

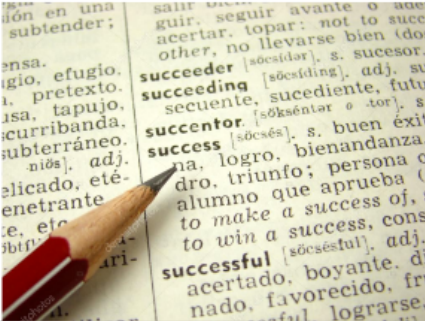
Basadas en otro tipo de inferencia lingüística

- etiquetado gramatical (categorías gramaticales: sustantivo, verbo, etc)

Contexto de etiquetado

- etiqueta anterior y siguiente

Identificación basada en palabras

| Palabra actual ¿Está en mi diccionario? | Palabra previa o siguiente |
|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | <ul style="list-style-type: none"> • “a las” => hora • “en” => lugar • “at...” => lugar (inglés) |

Substring de una palabra

| Substring | Categoría | Ejemplo |
|-----------|-----------|--------------------------------------------------------------------------------------------------------------|
| oxa | Drogas | Cotrimoxazole |
| field | Lugares | Wethersfield Banfield |
| : | Películas | El señor de los anillos: El retorno del rey 2001: Odisea del espacio South Park: Bigger Longer & Uncut |

Forma de una palabra

| Forma | Nombre de entidad |
|--------|-------------------|
| Xx-xxx | Varicela-zóster |
| xXXX | mRNA |
| XXXd | CPA1 |

Identificación basada otro tipo de inferencia lingüística

Adjetivo + Sustantivo + Sustantivo => Nombre de entidad

Identificación basada en contexto de etiquetado

PERSONA + PERSONA + ????? => Posiblemente sea PERSONA

Juan Carlos **Perez**
P P ?

Algoritmos de inferencia

- Greedy Inference
- Beam Inference
- Viterbi Inference
- CRFs