Extracción de la información

Extracción de la información

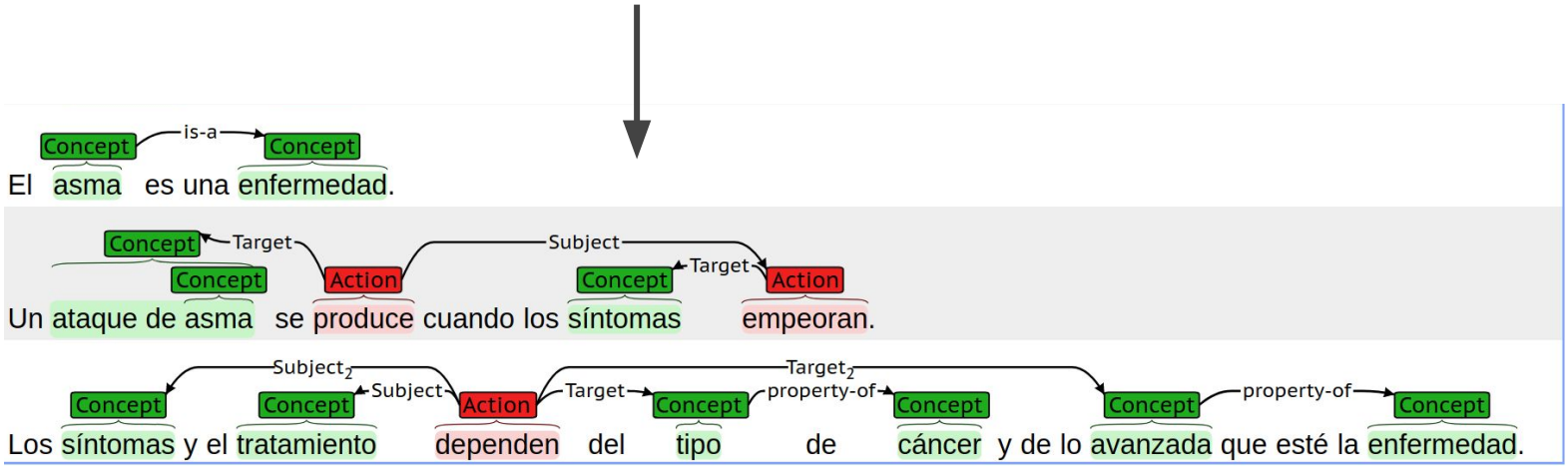
Transformar información no estructurada a información estructurada

El asma es una enfermedad. Un ataque de asma se produce cuando los síntomas empeoran. Los síntomas y el tratamiento dependen del tipo de cáncer y de lo avanzada que esté la enfermedad.

Extracción de la información

Transformar información no estructurada a información estructurada

El asma es una enfermedad. Un ataque de asma se produce cuando los síntomas empeoran. Los síntomas y el tratamiento dependen del tipo de cáncer y de lo avanzada que esté la enfermedad.



Extracción de la información

- Etiquetado de entidades: **Named entity recognition**
  - Reconocer organizaciones, lugares, ubicaciones, etc. (Ej: "New York Times", "Jorge Newbery")

Extracción de la información

- Etiquetado de entidades: **Named entity recognition**
  - Reconocer organizaciones, lugares, ubicaciones, etc. (Ej: "New York Times", "Jorge Newbery")
- Detección de relaciones: **Relation extraction**
  - Detectar la relación entre entidades dentro del texto (Ej: "es dueño de..", "es padre de…")

Extracción de la información

- Etiquetado de entidades: **Named entity recognition**
  - Reconocer organizaciones, lugares, ubicaciones, etc. (Ej: "New York Times", "Jorge Newbery")
- Detección de relaciones: **Relation extraction**
  - Detectar la relación entre entidades dentro del texto (Ej: "es dueño de..", "es padre de…")
- Detección de eventos: **Event extraction**
  - Detectar eventos dependiendo del dominio (Ej: se reservó un ticket de avión a tal hora)

Extracción de la información

- Etiquetado de entidades: **Named entity recognition**
  - Reconocer organizaciones, lugares, ubicaciones, etc. (Ej: "New York Times", "Jorge Newbery")
- Detección de relaciones: **Relation extraction**
  - Detectar la relación entre entidades dentro del texto (Ej: "es dueño de..", "es padre de...")
- Detección de eventos: **Event extraction**
  - Detectar eventos dependiendo del dominio (Ej: se reservó un ticket de avión a tal hora)
- Detección de expresiones temporales: **Temporal expression**
  - Poder ubicar en un calendario los eventos que se nombran en el texto (Ej: "el 2/3 a las 14hs")

Extracción de la información

- Etiquetado de entidades: **Named entity recognition**
  - Reconocer organizaciones, lugares, ubicaciones, etc. (Ej: "New York Times", "Jorge Newbery")
- Detección de relaciones: **Relation extraction**
  - Detectar la relación entre entidades dentro del texto (Ej: "es dueño de..", "es padre de...")
- Detección de eventos: **Event extraction**
  - Detectar eventos dependiendo del dominio (Ej: se reservó un ticket de avión a tal hora)
- Detección de expresiones temporales: **Temporal expression**
  - Poder ubicar en un calendario los eventos que se nombran en el texto (Ej: "el 2/3 a las 14hs")
- Detección de estructura: **Template filling**
  - Obtención de información relevante a un esquema.

Extracción de la información

Named Entity Recognition

Entidades: Cualquier cosa que pueda ser referenciada con un **NOMBRE PROPIO** (a veces se extiende esta definición con etiquetas temporales, cantidades o precios)

Extracción de la información

Named Entity Recognition

Entidades: Cualquier cosa que pueda ser referenciada con un **NOMBRE PROPIO** (a veces se extiende esta definición con etiquetas temporales, cantidades o precios)

Categorías más populares:

- PER (personas, personajes)
- ORG (compañías, equipos de fútbol, ongs, etc)
- LOC (regiones, montañas, mar, etc)
- GPE (ciudades, provincias, países, etc)
- FAC (puentes, edificios, aeropuertos, etc)
- VEH (autos, colectivos, aviones, etc)
- etc

Extracción de la información

Named Entity Recognition

Entidades: Cualquier cosa que pueda ser referenciada con un **NOMBRE PROPIO** (a veces se extiende esta definición con etiquetas temporales, cantidades o precios)

Categorías más populares:

- PER (personas, personajes)
- ORG (compañías, equipos de fútbol, ongs, etc)
- LOC (regiones, montañas, mar, etc)
- GPE (ciudades, provincias, países, etc)
- FAC (puentes, edificios, aeropuertos, etc)
- VEH (autos, colectivos, aviones, etc)
- etc

A veces sólo interesa encontrar las entidades y no su categoría, ejemplo:

```
<ENTITY url="https://en.wikipedia.org/wiki/Michael_Jordan"> Michael Jordan </ENTITY> es un jugador profesional en los <ENTITY url="http://en.wikipedia.org/wiki/Chicago_Bulls"> Chicago Bulls  </ENTITY>
```

Extracción de la información

Named Entity Recognition

Dificultad:

- ¿**Qué es** una entidad?
- ¿**Qué no es** una entidad?

Extracción de la información

Named Entity Recognition

Dificultad:

- ¿**Qué es** una entidad?
- ¿**Qué no es** una entidad?
- ¿Cómo **desambiguar**?
  - Ej: "***Jorge Newbery***"

Aeropuerto

Club

Persona

Estación de tren

Extracción de la información

Named Entity Recognition

Dificultad:

- ¿**Qué es** una entidad?
- ¿**Qué no es** una entidad?
- ¿Cómo **desambiguar**?
  - Ej: "***Jorge Newbery***"
- ¿Cuáles son los **límites** de la entidad?
  - Ej: "***New York Times es un*** "

`<GPE> New York </GPE>`

`<ORG> New York Times </ORG>`

Extracción de la información

Named Entity Recognition

Enfoque: **aprendizaje supervisado**

“San Antonio Spurs enfrenta una dura serie de playoffs ante Golden State Warriors: pierde 2-0 y mañana, desde las 22.30, en el AT&T Center, buscará comenzar con la remontada. “

1) Etiquetado manual

Palabra	Etiqueta
San	ORG
Antonio	ORG
Spurs	ORG
enfrenta	X
...	...

2) Extracción de Features (por palabra)

- ➔ POS tagging de w_i
- ➔ Forma de la palabra w_i
- ➔ Presencia de w_i en algún listado (RAE, gazetteer)
- ➔ w_i contiene prefijo x (para todo |x| <= 4)
- ➔ w_i contiene sufijo x (para todo |x| <= 4)
- ➔ w_i está en mayúscula
- ➔ Presencia de apóstrofos.
- ➔ Etc

Extracción de la información

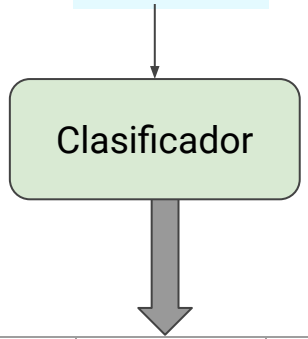
Named Entity Recognition

1) Etiquetado **2) Extracción de Features** **3) Entrenamiento de un clasificador**

San	*Antonio*	*Spurs*	*enfrenta*	*una*	*dura*	*serie*
FEATURES						
NNP Xxx enRae: SI	NNP Xxxxxxx enRae: SI	NNP Xxxx enRae: NO	VB xxxxxxxx enRae: SI	DT xxx enRae: SI	ADV Xxxx enRae: SI	N xxxxx enRae: SI

VERSIÓN 1

Un clasificador convencional.



TAGS					
ORG	**ORG**	**ORG**	**X**	???	???

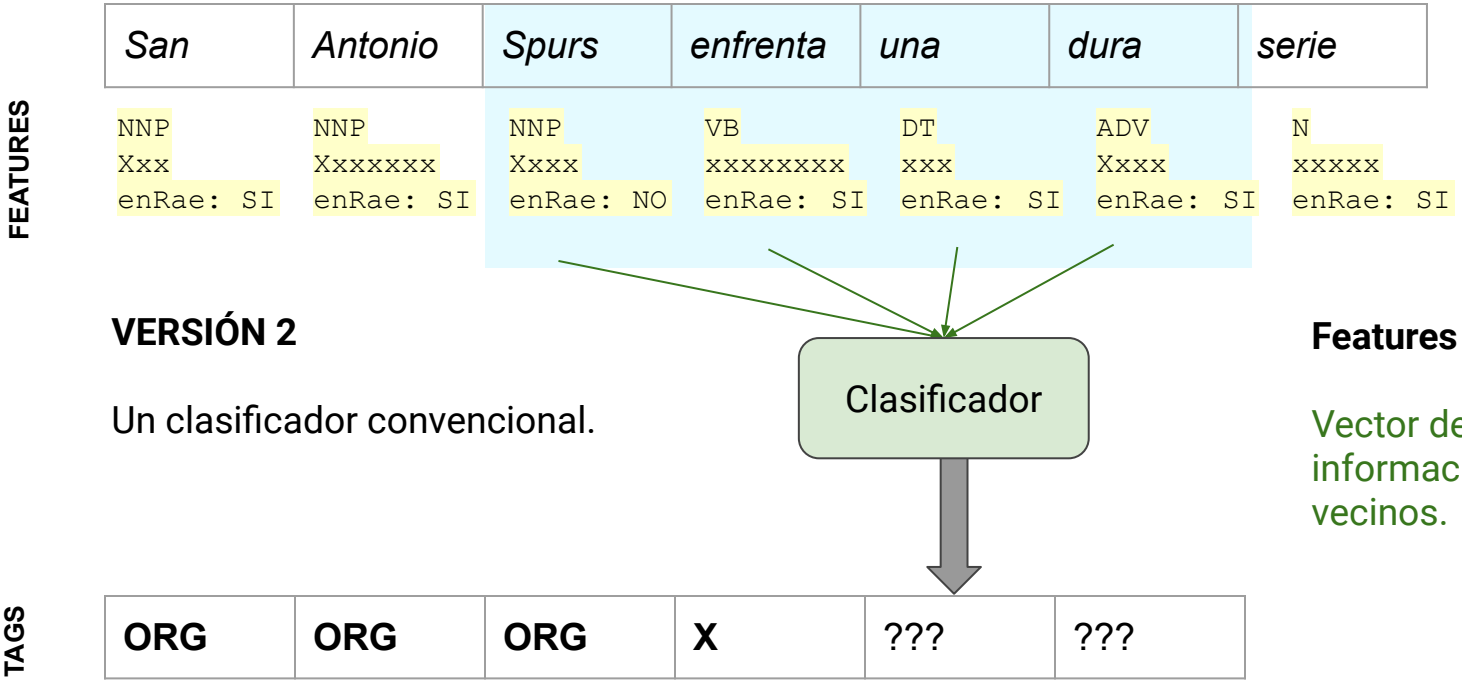
PROBLEMA

Muy pocos ejemplos en entrenamiento de cada caso particular: La información relevante suele estar en el **contexto**!

Extracción de la información

Named Entity Recognition

1) Etiquetado **2) Extracción de Features** **3) Entrenamiento de un clasificador**



VERSIÓN 2

Un clasificador convencional.

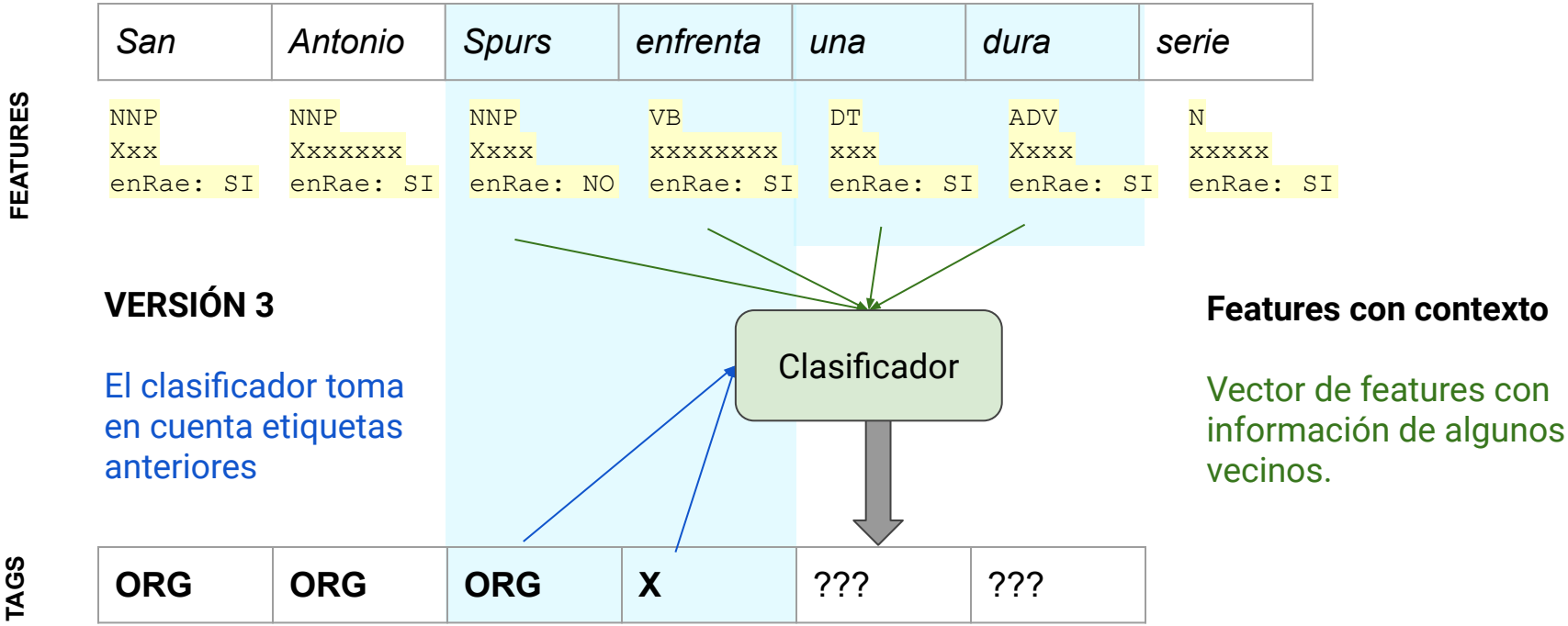
Features con contexto

Vector de features con información de algunos vecinos.

Extracción de la información

Named Entity Recognition

1) Etiquetado **2) Extracción de Features** **3) Entrenamiento de un clasificador**



Extracción de la información

Named Entity Recognition

1) Etiquetado **2) Extracción de Features** **3) Entrenamiento de un clasificador**