

Extracción de información

INTRO - NER

Extracción de información

Entidades y relaciones pre-especificadas:

Ejemplo:

- **Empresas y directivos** (u **organismos de gobierno y legisladores, políticos**)
 - quién ocupa
 - qué cargo
 - cuándo

El director General de Secretaría del Ministerio del Interior, Charles Carrera le pidió la renuncia al subdirector del Instituto Nacional de Rehabilitación (INR), Gustavo Belarra por no tener el diploma de sociólogo y haber incluido en su currículum que ya tenía el título.

Fuentes del INR dijeron a El País que Belarra presentó su tesis y no la defendió, con lo cual no finalizó su carrera.

La doctora Ana Gabriela González Gargano será quien reemplace al sociólogo Jorge Papadópolos en la dirección General de Secretaría del Ministerio de Educación y Cultura.

La razón por la que Papadópolos fue removido habría sido una discusión que mantuvo con otra de las autoridades de la cartera.

El director General de Secretaría del Ministerio del Interior, **Charles Carrera** le pidió la renuncia al subdirector del Instituto Nacional de Rehabilitación (INR), **Gustavo Belarra** por no tener el diploma de sociólogo y haber incluido en su currículum que ya tenía el título.

Fuentes del INR dijeron a El País que **Belarra** presentó su tesis y no la defendió, con lo cual no finalizó su carrera.

La doctora **Ana Gabriela González Gargano** será quien reemplace al sociólogo **Jorge Papadópolos** en la dirección General de Secretaría del Ministerio de Educación y Cultura.

La razón por la que **Papadópolos** fue removido habría sido una discusión que mantuvo con otra de las autoridades de la cartera.

El director General de **Secretaría del Ministerio del Interior**, **Charles Carrera** le pidió la renuncia al subdirector del **Instituto Nacional de Rehabilitación (INR)**, **Gustavo Belarra** por no tener el diploma de sociólogo y haber incluido en su currículum que ya tenía el título.

Fuentes del **INR** dijeron a **El País** que **Belarra** presentó su tesis y no la defendió, con lo cual no finalizó su carrera.

La doctora **Ana Gabriela González Gargano** será quien reemplace al sociólogo **Jorge Papadópolos** en la dirección General de **Secretaría del Ministerio de Educación y Cultura**.

La razón por la que **Papadópolos** fue removido habría sido una discusión que mantuvo con otra de las autoridades de la cartera.

El director General de Secretaría del Ministerio del Interior, Charles Carrera le pidió la renuncia al subdirector del Instituto Nacional de Rehabilitación (INR), Gustavo Belarra por no tener el diploma de sociólogo y haber incluido en su currículum que ya tenía el título.

Fuentes del INR dijeron a El País que Belarra presentó su tesis y no la defendió, con lo cual no finalizó su carrera.

La doctora Ana Gabriela González Gargano será quien reemplace al sociólogo Jorge Papadópulos en la dirección General de Secretaría del Ministerio de Educación y Cultura.

La razón por la que Papadópulos fue removido habría sido una discusión que mantuvo con otra de las autoridades de la cartera.

El director General de Secretaría del Ministerio del Interior, Charles Carrera le pidió la renuncia al subdirector del Instituto Nacional de Rehabilitación (INR), Gustavo Belarra por no tener el diploma de sociólogo y haber incluido en su currículum que ya tenía el título.

Fuentes del INR dijeron a El País que Belarra presentó su tesis y no la defendió, con lo cual no finalizó su carrera.

La doctora Ana Gabriela González Gargano será quien reemplace al sociólogo Jorge Papadópulos en la dirección General de Secretaría del Ministerio de Educación y Cultura.

La razón por la que Papadópulos fue removido habría sido una discusión que mantuvo con otra de las autoridades de la cartera.

El director General de Secretaría del Ministerio del Interior, Charles Carrera le pidió la renuncia al subdirector del Instituto Nacional de Rehabilitación (INR), Gustavo Belarra por no tener el diploma de sociólogo y haber incluido en su currículum que ya tenía el título.

Fuentes del INR dijeron a El País que Belarra presentó su tesis y no la defendió, con lo cual no finalizó su carrera.

La doctora Ana Gabriela González Gargano será quien reemplace al sociólogo Jorge Papadópolos en la dirección General de Secretaría del Ministerio de Educación y Cultura.

La razón por la que Papadópolos fue removido habría sido una discusión que mantuvo con otra de las autoridades de la cartera.

Extracción de información

Persona	Cargo	Organización	Tipo	Fecha
Gustavo Belarra	Subdirector	INR	Cese	Publicación ??
Ana Gabriela González	Director General de Secretaría	MEC	Ingreso	Publicación ??
Jorge Papadopoulos	Director General de Secretaría	MEC	Cese	publicación

Proceso de información no estructurada - generando datos

Definición IE

- Transforma Información transmitida por texto en información estructurada (BD)
- Objetivos
 - Presentar la información en formato compacto y eventualmente validado
 - Acumular datos en Bases de Datos (Bases de Conocimientos)
 - Estas bases son requeridas por sistemas de minería, eventualmente sistemas NLP que requieren conocimiento para realizar inferencias

¿Para qué EI ?

- Una empresa quiere seguir las reacciones sobre un nuevo lanzamiento de un producto en blogs web.
- Una compañía quiere usar las noticias que recibe de una agencia de prensa para construir una descripción detallada de todas las tendencias tecnológicas en el desarrollo de tecnologías de semiconductores. La compañía también quiere un registro de todas las transacciones comerciales involucradas en este desarrollo.
- Una agencia espacial permite a los astronautas consultar grandes cantidades de documentación técnica mediante lenguaje natural.
- Un gobierno está recopilando datos sobre un desastre natural y quiere transmitir de modo urgente a los servicios de emergencia un resumen de los últimos datos disponibles.
- Un profesional o académico del área legal está interesado en estudiar las decisiones de los jueces en los acuerdos de divorcio y los criterios subyacentes. Tiene miles de decisiones judiciales a su disposición.
- Un grupo de investigación biomédica está investigando un nuevo tratamiento y quiere conocer todas las formas posibles en que un grupo específico de proteínas puede interactuar con otras proteínas y cuáles son los resultados exactos de estas interacciones. Hay decenas de miles de artículos e informes técnicos para estudiar.

En **América Latina**, cuyos aeropuertos y aerolíneas emplean a unas **430.000** personas, “el tráfico aéreo prácticamente ha desaparecido”, le dijo a **BBC Mundo Cristina Fernández**, directora de la división para **América Latina y el Caribe** del **Consejo Internacional de Aeropuertos (ACI-LAC)**, el **3 de mayo** en **Atlanta**.

Personas

Lugares (GPE)

Organización

Número

fecha

Objetos que se identifican con un nombre.

En general expresiones multipalabra

Se incluyen números, fechas, cantidades de diverso tipo.

La intención es asociar unívocamente las entidades con entes del mundo, pero no siempre es posible sin procesos posteriores.

Entidades y clases

Entidades con nombre (*Named entities*)

- Cosas identificables, discretas, en principio con identificación con nombre en su clase
 - Países, ciudades, ríos
 - Personas
 - Organizaciones
 - Una casa, un martillo, una silla, no entrarían en la categoría (hay excepciones, casas con nombre)
- Se agregaron otras unidades de texto que interesaba identificar y que no son entidades con nombre
 - Fechas
 - Números que denotan cantidad, medidas con sus unidades
 - Porcentajes
 - Genes, proteínas
 - Compuestos químicos
 - Enfermedades, medicamentos

¿Para qué se reconocen ?

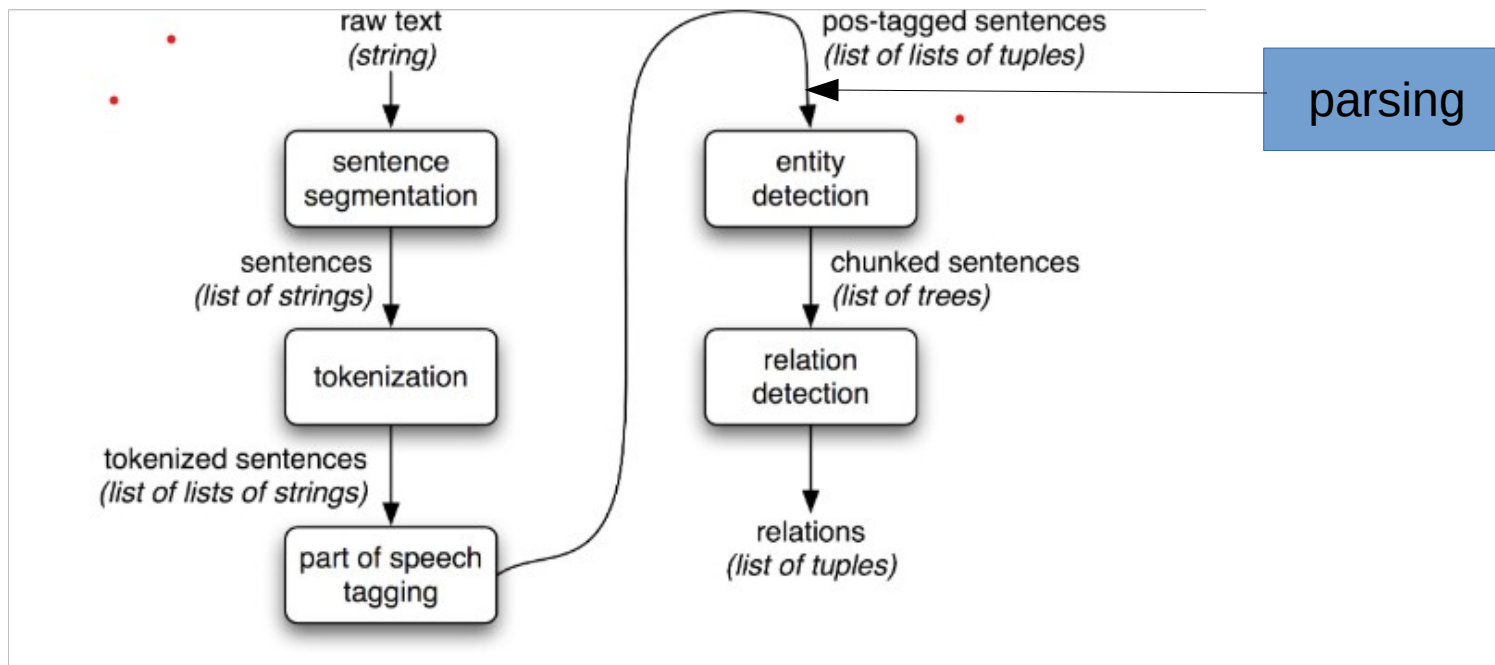
- ♦ El trabajo en IE con MUC mostró gran parte de las relaciones se daban entre entidades con nombre.
- ♦ El análisis de sentimientos puede referir a políticos, empresas, productos, libros, películas : todos ellos son entidades con nombre
- ♦ Los dominios técnicos definen y nombran los conceptos esenciales, es útil extraer estos nombres para la extracción de información

Reconocimiento y clasificación

- Se trata de un problema secuencial: dado un texto se debe identificar subsecuencias de tokens que sean una entidad con nombre.
- Junto con la identificación se asigna una clase.

Ana_López_de_la_BBC entrevistó a el secretario de la ONU el 5_de_mayo en París .
ana_lópez_de_la_bbc entrevistar a el secretario de el onu el [?:?:5/5/?:?:?:?:?:?] en paris :
NP00SP0 VMIS3S0 SP DA0MS0 NCMS000 SP DA0FS0 NP00O00 DA0MS0 W SP NP00G00 Fp

Pipeline EI



NLTK book, chap 7

Métodos para reconocimiento de entidades con nombre

- Patrones o reglas manuales
- Aprendizaje supervisado
- Aprendizaje semisupervisado

Extracción de entidades con nombre, reglas manuales

- Un método directo para encontrar nombres propios es escribir un conjunto de patrones (con expresiones regulares):
- Título [token-mayúscula +] → persona siendo Título uno de {Sr., Sra., Ing., ...}
- Es posible también manejar una lista de nombres de pila, lo que permitirá resolver muchos casos.

Extracción de entidades con nombre, reglas manuales

- Criterios similares pueden encontrarse para organizaciones: tokens como SRL, SA, etc. identifican empresas.
- Palabras como Ministerio, Secretaría, Administración, ... identifican organizaciones gubernamentales
- Es común tener largas listas de personas, lugares geográficos, etc en los sistemas manuales.

Extracción de entidades con nombre, reglas manuales

- Las listas funcionan bien, pero deben ser actualizadas periódicamente.
- En general, un sistema manual basado en patterns y listas sigue siendo una solución razonable para algunos tipos de entidades con nombre.
- Hay casos que requieren desambiguar (p.ej., Uruguay, país y río). Difícil solo con listas y *patterns*.
- Es bueno contar con un corpus para la evaluación.

Extracción de entidades con nombre, aprendizaje supervisado

- Requiere un corpus con suficientes ejemplos para los distintos casos de salida.
- Es un problema secuencial, es usual modelarlo como etiquetado BIO.

Fuentes del INR dijeron a El País que Belarra presentó su
O O B O O B I O B O O

Extracción de entidades con nombre, aprendizaje supervisado

- Esquema BIO
 - B - 1^{er} token del segmento (o único)
 - I – tokens siguientes del segmento
 - O – tokens externos

Fuentes del INR dijeron a El País que Belarra presentó su
O O B O O B I O B O O

Extracción de entidades con nombre, aprendizaje supervisado

Fuentes del INR dijeron a El País que Belarra presentó su

O O B-o O O B-o I-o O B-p O O

De hecho, hay que tener varios tipos de token B e I.

- B-o Comienzo de organización
- I-o Continuación de organización

- B-p Comienzo de persona
- I-p Interior de persona


Idem para otras clases

Extracción de entidades con nombre, aprendizaje supervisado

- Habría que tener $2n+1$ estados, para n clases distintas.
- Características de los patrones (p.ej., Sr. precediendo un nombre) se pueden capturar por atributos (en MEMM o CRF). En HMM hay que usar más estados.
- Se ha propuesto también un esquema BLOU en lugar de BIO, empíricamente funcionó mejor (1%)
 - B- comienzo
 - I – interior
 - L – último
 - U – unitario
 - O - externo

Desambiguación en Wikimedia

Cristina Fernández (desambiguación)

página de desambiguación de Wikimedia / De Wikipedia, la enciclopedia libre 

15 de octubre

17 de septiembre

4 de septiembre

...

Cristina Fernández puede referirse a:

- **Cristina Fernández** (siglo XI), condesa asturiana, madre de doña Jimena Díaz (1046-1116, esposa del Cid).
- **Cristina Fernández Cubas** (1945—), escritora española.
- **Cristina Fernández** (1946—), cantautora uruguaya, del dúo de música popular Washington Carrasco y Cristina Fernández.
- **Cristina Fernández de Kirchner** (1953—), política, actual vicepresidenta de la Nación Argentina, expresidenta de la República Argentina (desde 2007-2015) y senadora (2017-2019).
- **Cristina Fernández (música)** ,música popular uruguaya.
- **Cristina Fernández (actriz)** , actriz argentina.

Otros aspectos

- Anáforas, pronombres
- Intersectar información de distintos documentos, particularmente si realizamos agregación de información.
- Desambiguar referencias por nombres iguales
- Evaluación : se han usado las medidas estándar de IR, con match exacto. Esto es un problema: un acierto parcial cuenta como 2 errores (un falso positivo y un falso negativo).
- Se ha reportado hasta un 90% de medida F,
- Últimamente con RN y el modelo Bert se reportó mejor F.