



Modelos de Inteligencia Artificial

Curso de Especialización de Inteligencia Artificial y Big Data
IES Gran Capitán 2024/25

UNIDAD 3.

Procesamiento del Lenguaje Natural y sus aplicaciones. Potencial y limitaciones.

Índice de contenidos

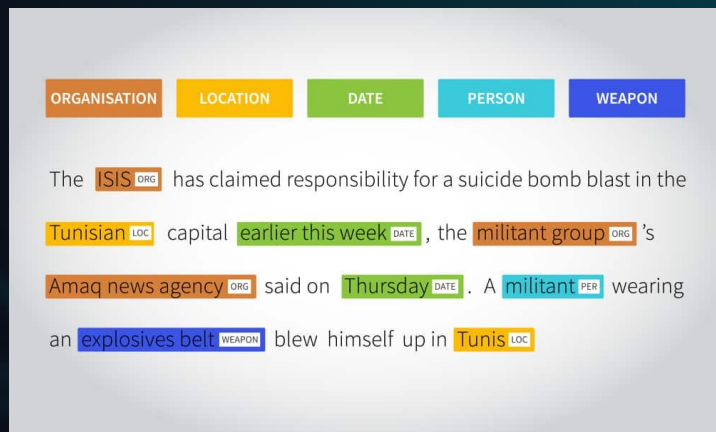
1. Procesamiento de lenguaje natural
2. Potencial de las técnicas existentes de procesamiento de lenguaje. Limitaciones
 1. Potencial de procesamiento del lenguaje natural
 1. Reconocimiento del habla (ASR, Automatic Speech Recognition)
 2. Síntesis de texto o voz
 3. Detección de entidades nombradas (NER, Named Entity Recognition)
 4. Traducción automática
 5. Similitud de textos
 6. Análisis del sentimiento
 2. Modelos y técnicas existentes
 1. BERT (Bidirectional Encoder Representations from Transformers)
 2. Otros modelos
 3. Limitaciones. La ambigüedad
3. Formación del investigador en PLN.
4. Elaboración de un sistema de procesamiento de lenguaje orientado a una tarea específica



2. Potencial de las técnicas existentes de procesamiento de lenguaje. Limitaciones.

2.1. Potencial de procesamiento del lenguaje natural

Detección de entidades nombradas (NER, named entity recognition)



The screenshot displays a text processing interface for Named Entity Recognition. At the top, there are five colored buttons representing entity categories: ORGANISATION (orange), LOCATION (yellow), DATE (green), PERSON (light blue), and WEAPON (dark blue). Below these buttons, a text snippet is shown with various words highlighted in colored boxes corresponding to these categories. The text is: "The **ISIS** has claimed responsibility for a suicide bomb blast in the **Tunisian** capital **earlier this week**, the **militant group**'s **Amaq news agency** said on **Thursday**. A **militant** wearing an **explosives belt** blew himself up in **Tunis**".

Entity	Category
ISIS	ORGANISATION
Tunisian	LOCATION
earlier this week	DATE
militant group	ORGANISATION
Amaq news agency	ORGANISATION
Thursday	DATE
militant	PERSON
explosives belt	WEAPON
Tunis	LOCATION

Consiste en trocear el texto (tokenizar) de tal manera que se detecten las palabras clave. Se trata de una labor de extracción de información que localiza y clasifica en categorías (normalmente personas, organizaciones, lugares y cantidades) las entidades encontradas en un texto.



<https://es.shaip.com/blog/named-entity-recognition-and-its-types/>

¿Cómo se relaciona NER con PNL?

El PLN ayuda a desarrollar máquinas inteligentes capaces de extraer significado del habla y el texto. Machine Learning ayuda a estos sistemas inteligentes a seguir aprendiendo entrenando grandes cantidades de lenguaje natural conjuntos de datos.

En general, la PNL consta de tres categorías principales:

- Entender la estructura y las reglas del lenguaje - **Sintaxis**
- Deducir el significado de las palabras, el texto y el habla e identificar sus relaciones.

Semántica

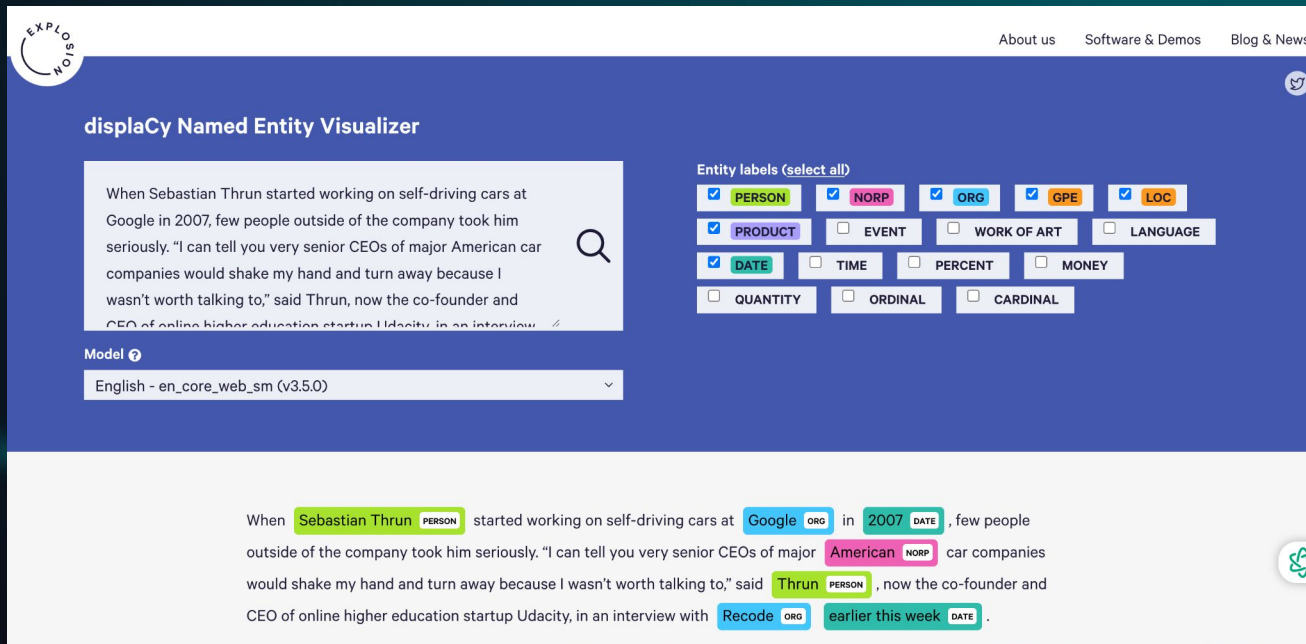
- Identificar y reconocer palabras habladas y transformarlas en texto. - **Habla**

NER ayuda en la parte semántica de la PNL, extrayendo el significado de las palabras, identificándolas y localizándolas en función de sus relaciones.

Ejemplos detección de entidades nombradas

Apple ORG today DATE announced the
second QUANTITY generation iPhone SE COMM
a powerful new iPhone COMM featuring
a 4.7- inch QUANTITY Retina HD display.

Ejemplos detección de entidades nombradas



The screenshot displays the 'displaCy Named Entity Visualizer' interface. At the top left is the 'EXPLOSION' logo. The top right contains links for 'About us', 'Software & Demos', and 'Blog & News'. The main title 'displaCy Named Entity Visualizer' is on the left. Below it is a text input area containing a paragraph about Sebastian Thrun. To the right of the input is a search icon. Below the input is a 'Model' dropdown menu set to 'English - en_core_web_sm (v3.5.0)'. On the right side, there is a section for 'Entity labels (select all)' with checkboxes for various categories: PERSON, NORP, ORG, GPE, LOC, PRODUCT, EVENT, WORK OF ART, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, and CARDINAL. The bottom section shows the output text with entities highlighted and labeled: 'Sebastian Thrun' (PERSON), 'Google' (ORG), '2007' (DATE), 'American' (NORP), 'Thrun' (PERSON), 'Recode' (ORG), and 'earlier this week' (DATE).

EXPLOSION

About us Software & Demos Blog & News

displaCy Named Entity Visualizer

When Sebastian Thrun started working on self-driving cars at Google in 2007, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun, now the co-founder and CEO of online higher education startup Udacity, in an interview.

Model

English - en_core_web_sm (v3.5.0)

Entity labels (select all)

☒ PERSON ☒ NORP ☒ ORG ☒ GPE ☒ LOC

☒ PRODUCT ☐ EVENT ☐ WORK OF ART ☐ LANGUAGE

☒ DATE ☐ TIME ☐ PERCENT ☐ MONEY

☐ QUANTITY ☐ ORDINAL ☐ CARDINAL

When Sebastian Thrun started working on self-driving cars at Google in 2007, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode earlier this week.



<https://demos.explosion.ai/displacy-ent>

Ejemplos detección de entidades nombradas

spaCy

Curso avanzado de spacy

```
pip install spacy
python -m spacy download es_core_news_sm
python -m spacy download en_core_news_sm
#Tras instalar SPacy y los modelos en español e inglés
import spacy
from spacy import displacy
from collections import Counter

import es_core_news_sm

nlp = es_core_news_sm.load()

doc = nlp('Carlos enseña a los alumnos a programar con NVIDIA')
print([(X.text, X.label_) for X in doc.ents])
```

Ejemplos detección de entidades nombradas

spaCy

Nota: mediante una orden gráfica y un entorno gráfico de iPython como Jupyter, es posible graficar los NER y la tokenización mediante `displacy.render(doc, style='dep', jupyter=True, options={'distance':120})`

La salida no gráfica es la siguiente:

```
[('Carlos', 'PER'), ('NVIDIA', 'ORG')]
```

Para saber más

En el siguiente recurso WEB se incluye material y ejemplos de NLTK y con Spacy:

- <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>

Para estudiar ejemplos de material mediante NeMo, se puede clicar en el siguiente link:

https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/token_classification.html