Introducción al Procesado de Lenguaje Natural

...

¿Qué es el NLP?

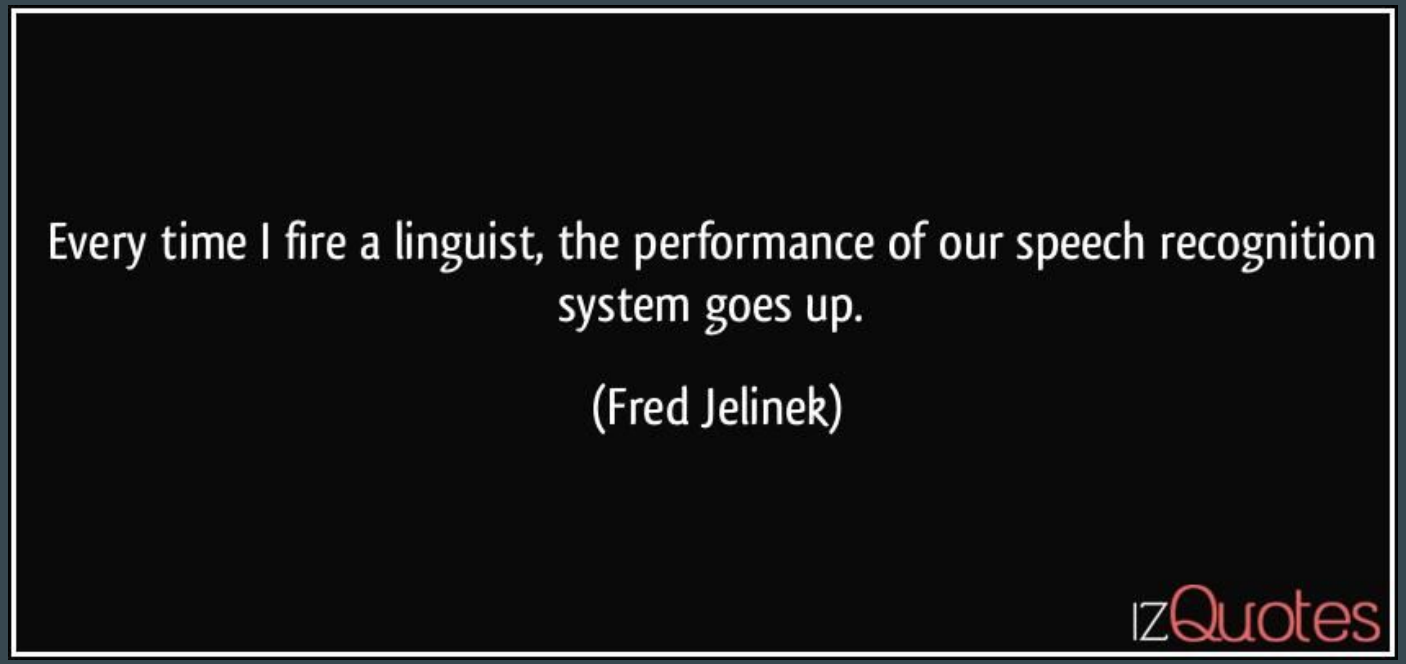
¿Qué es el NLP?

- Natural Language Processing (NLP)
- Intersección entre la IA, la lingüística y las ciencias de la computación
- Casi siempre se trabaja con texto, aunque también voz
- Es keyword matching, es normalización de textos, es ML, es clustering...
- Entendimiento humano-máquina

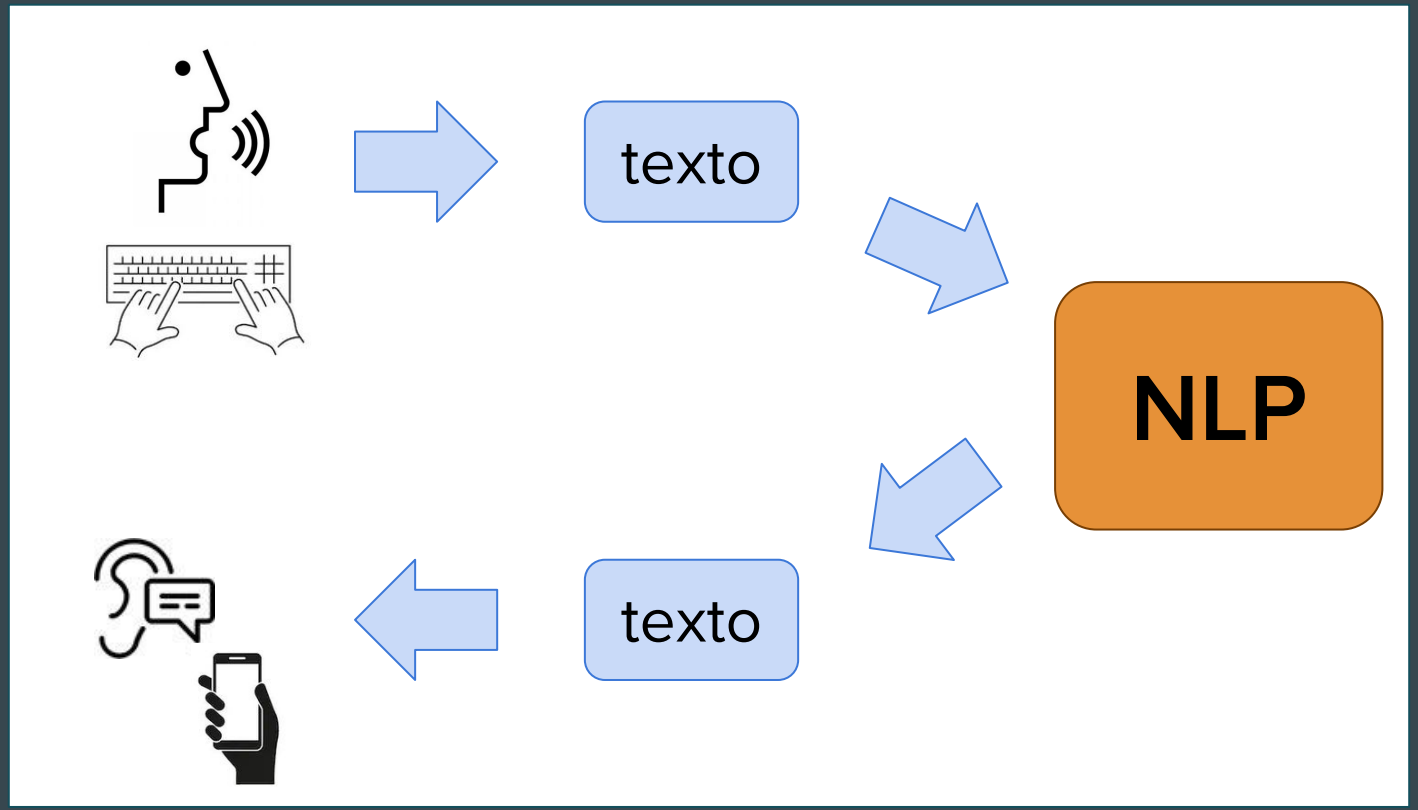
Imagen extraída de: https://aliz.ai/natural-language-processing-a-short-introduction-to-get-you-started/

Perfiles en NLP

Lingüistas, matemáticos, ingenieros, estadísticos, ...



Cadena típica en NLP

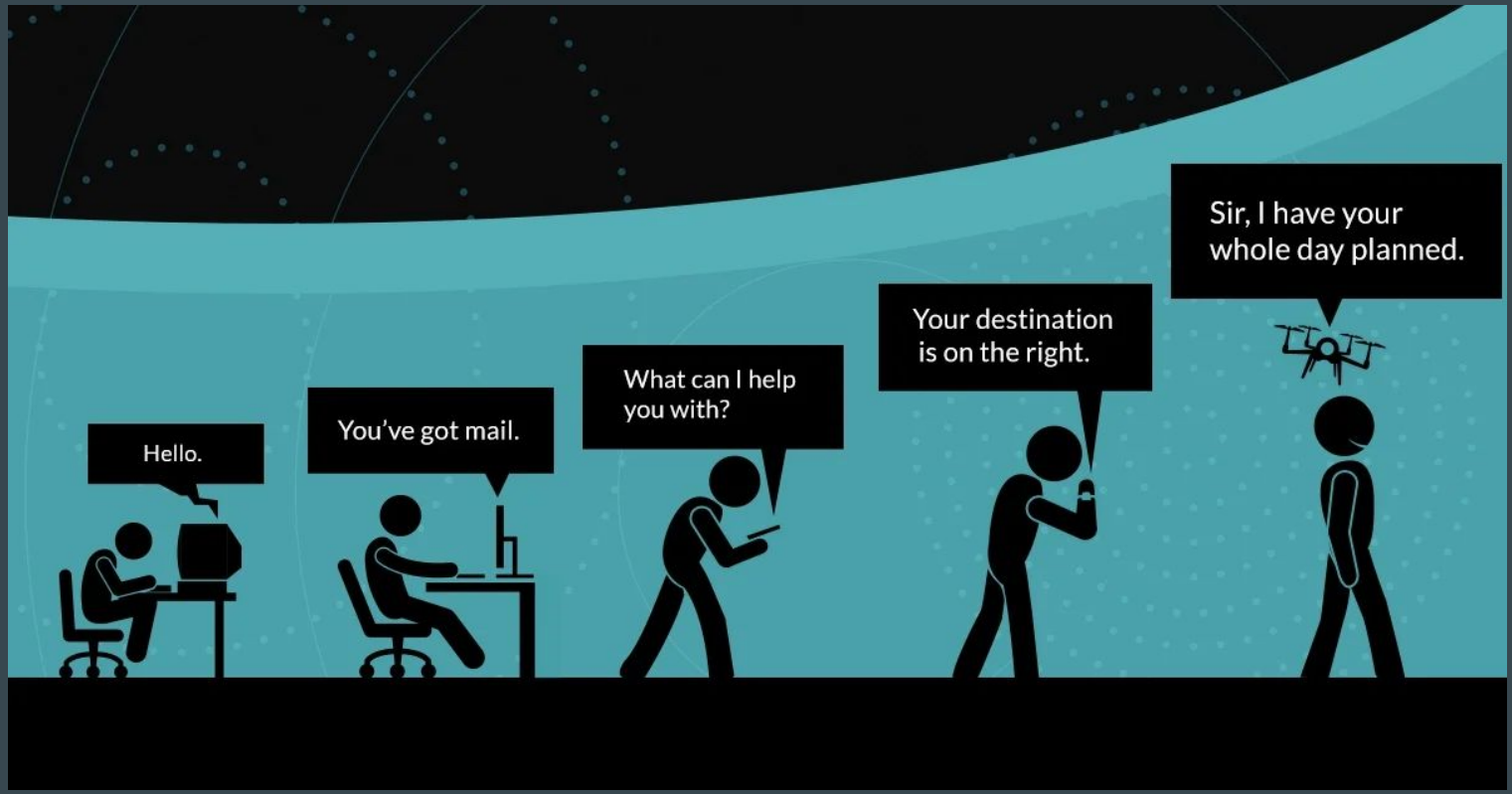


Tareas en NLP

Tareas en NLP

- Dominio concreto
  - Clasificación de documentos
  - Análisis de sentimiento en comentarios sobre marcas, productos, mercados, políticos, ...
  - Búsqueda de información en un buscador
  - Detección de los temas de moda en redes sociales
- End-to-end
  - Agentes conversacionales
  - Traducción bidireccional en tiempo real
  - Módulo de entendimiento de lenguaje (NLU) completo

Estado del arte



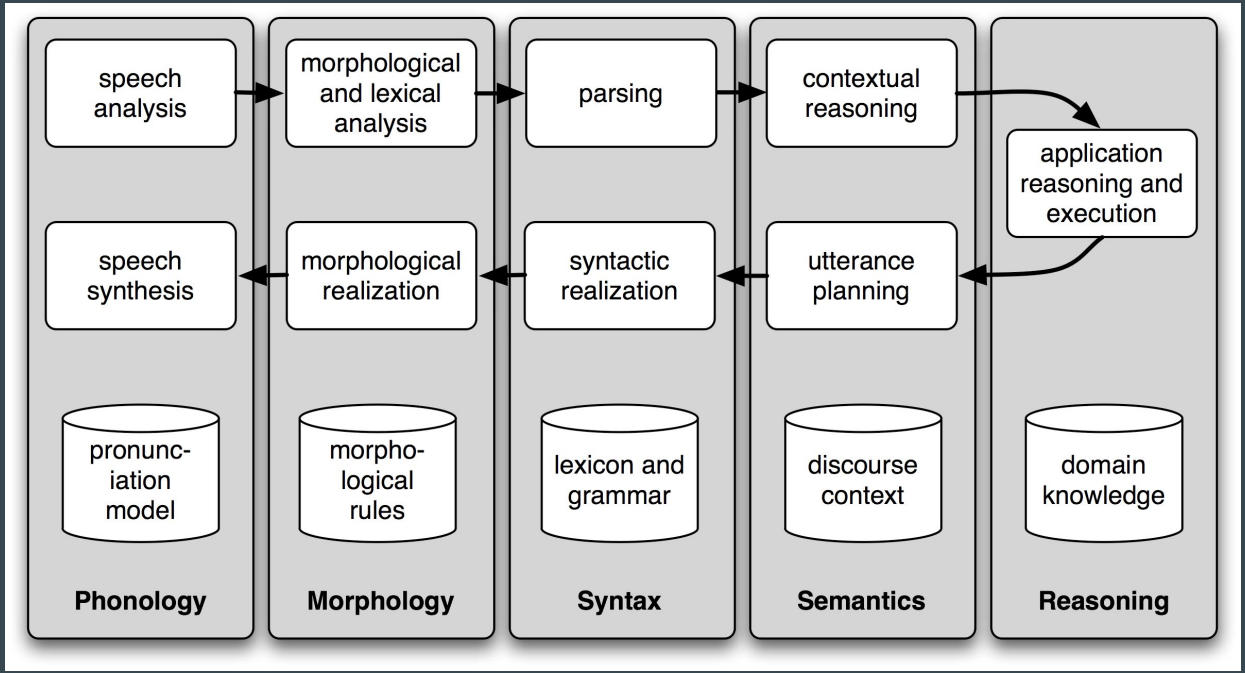
Estado del arte

Algunas tareas

Prácticamente Resueltas	• Detección de spam • PoS Tagging • “Named Entity Recognition”
Progresa adecuadamente	• Sentiment Analysis • Machine Translation • Desambiguación • Extraccion de Informacion
Aún muy lejos	• Diálogo • Parafrasear • Preguntas y Respuestas • Resumir texto • Detección de sarcasmo

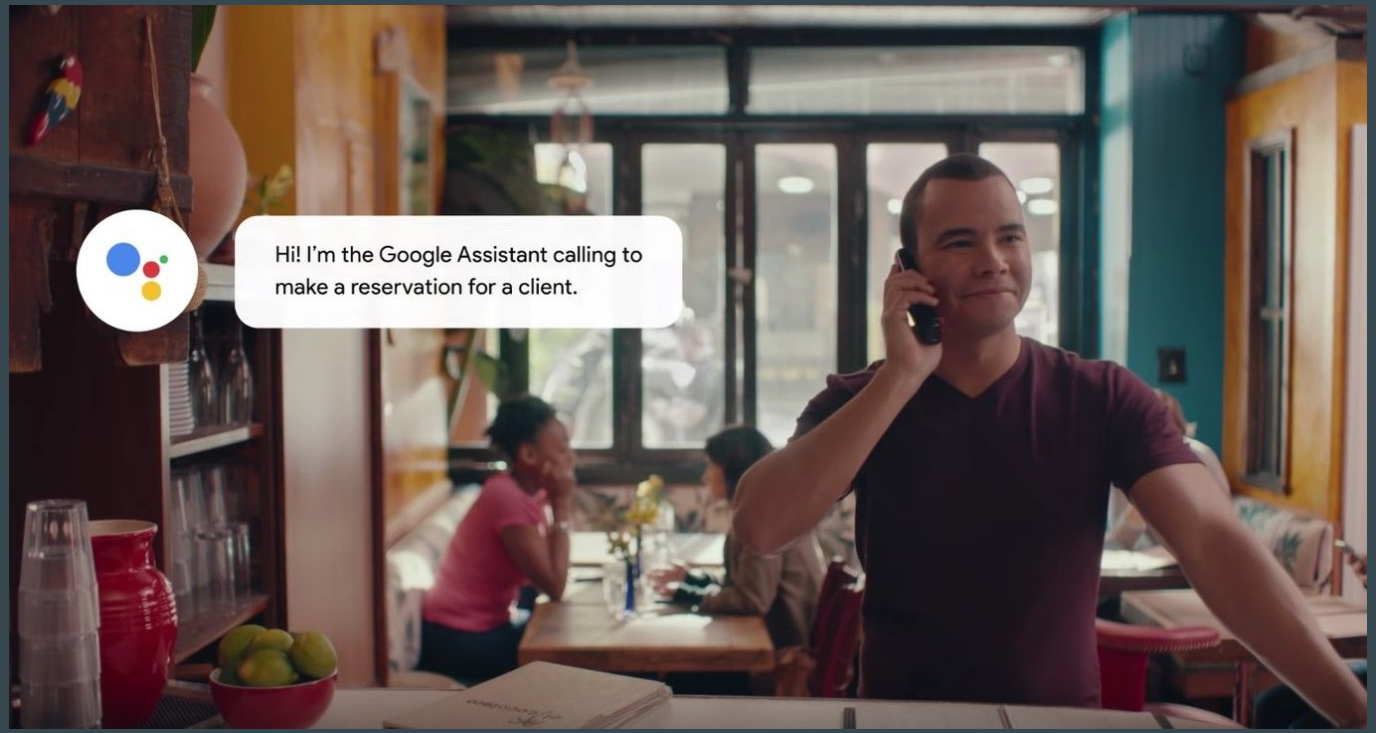
Tareas en NLP

- Algunas tareas de NLP en un asistente virtual



Estado del arte

Asistente de Google haciendo una reserva (demo)



Estado del arte

NLG con GPT-2, de OpenAI. Demo

SYSTEM PROMPT (HUMAN-WRITTEN)

Recycling is good for the world.

NO! YOU COULD NOT BE MORE WRONG!!

MODEL COMPLETION (MACHINE-WRITTEN, 25 TRIES)

Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a

Estado del arte

Query a Alexa. El día a día...

Procesado de Texto

Procesado de texto

¿Qué es un texto?

texto

Del lat. *textus;* propiamente 'trama', 'tejido'.

1. m. Enunciado o conjunto coherente de enunciados orales o escritos.

Procesado de texto

¿Qué es un texto?

texto

Del lat. *textus;* propiamente 'trama', 'tejido'.

1. m. Enunciado o conjunto coherente de enunciados orales o escritos.

coherencia

Del lat. *cohaerentia.*

1. f. Conexión, relación o unión de unas cosas con otras.

2. f. Actitud lógica y consecuente con los principios que se profesan.

3. f. *Fís.* **cohesión** (ll unión entre moléculas).

4. f. *Ling.* Estado de un sistema lingüístico o de un texto cuando sus componentes aparecen en conjuntos solidarios. *La coherencia del sistema de adverbios de lugar en español se manifiesta en tres grados.*

enunciado

De *enunciar.*

1. m. **enunciación.**

2. m. Secuencia de palabras delimitada por pausas muy marcadas, que puede estar constituida por una o varias oraciones.

3. m. *Ling.* Secuencia con valor comunicativo, sentido completo y entonación propia.

Procesado de texto

¿Qué es un texto?

texto

Del lat. *textus;* propiamente 'trama', 'tejido'.

1. m. Enunciado o conjunto coherente de enunciados orales o escritos.

coherencia

Del lat. *cohaerentia.*

1. f. Conexión, relación o unión de unas cosas con otras.

2. f. Actitud lógica y consecuente con los principios que se profesan.

3. f. *Fís.* **cohesión** (|| unión entre moléculas).

4. f. *Ling.* Estado de un sistema lingüístico o de un texto cuando sus componentes aparecen en conjuntos solidarios. *La coherencia del sistema de adverbios de lugar en español se manifiesta en tres grados.*

enunciado

De *enunciar.*

1. m. **enunciación.**

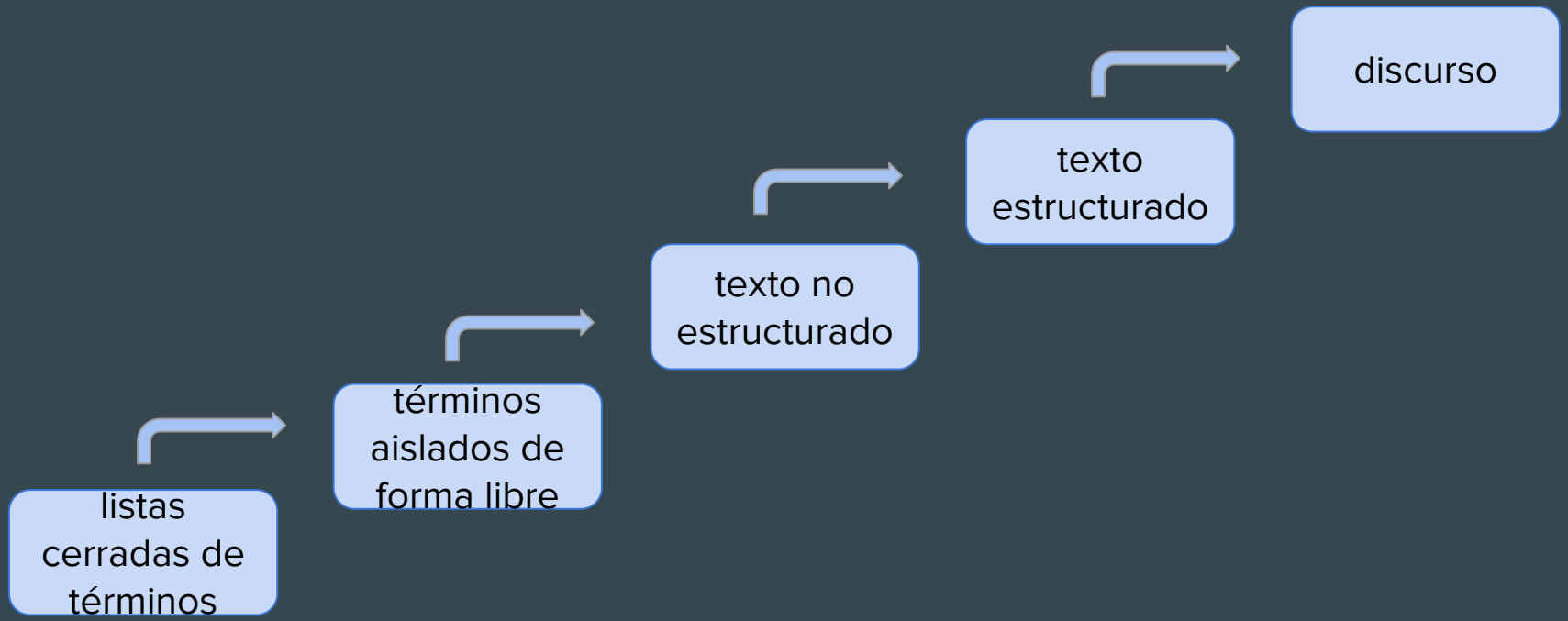
2. m. Secuencia de palabras delimitada por pausas muy marcadas, que puede estar constituida por una o varias oraciones.

3. m. *Ling.* Secuencia con valor comunicativo, sentido completo y entonación propia.

Un conjunto de símbolos que significan algo para alguien

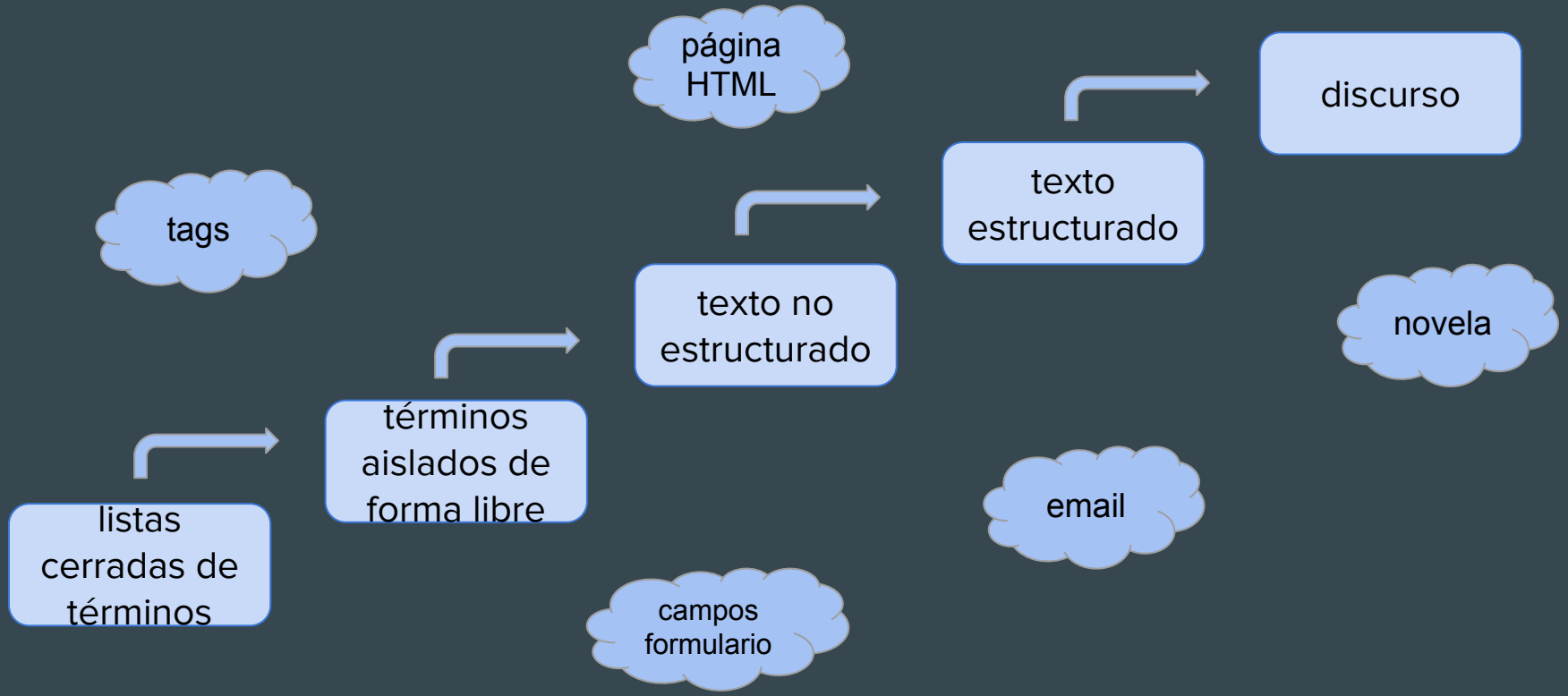
Procesado de texto

“Tipos” de texto



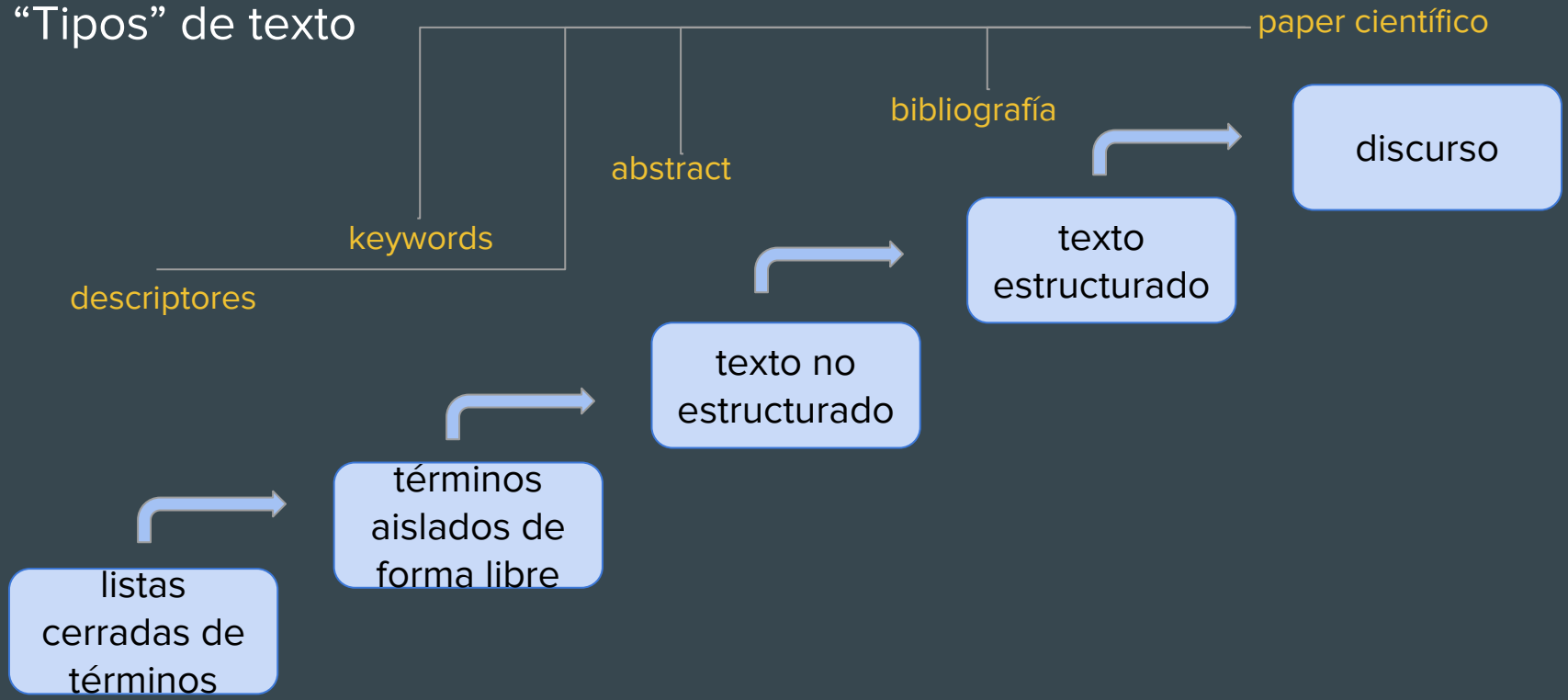
Procesado de texto

“Tipos” de texto



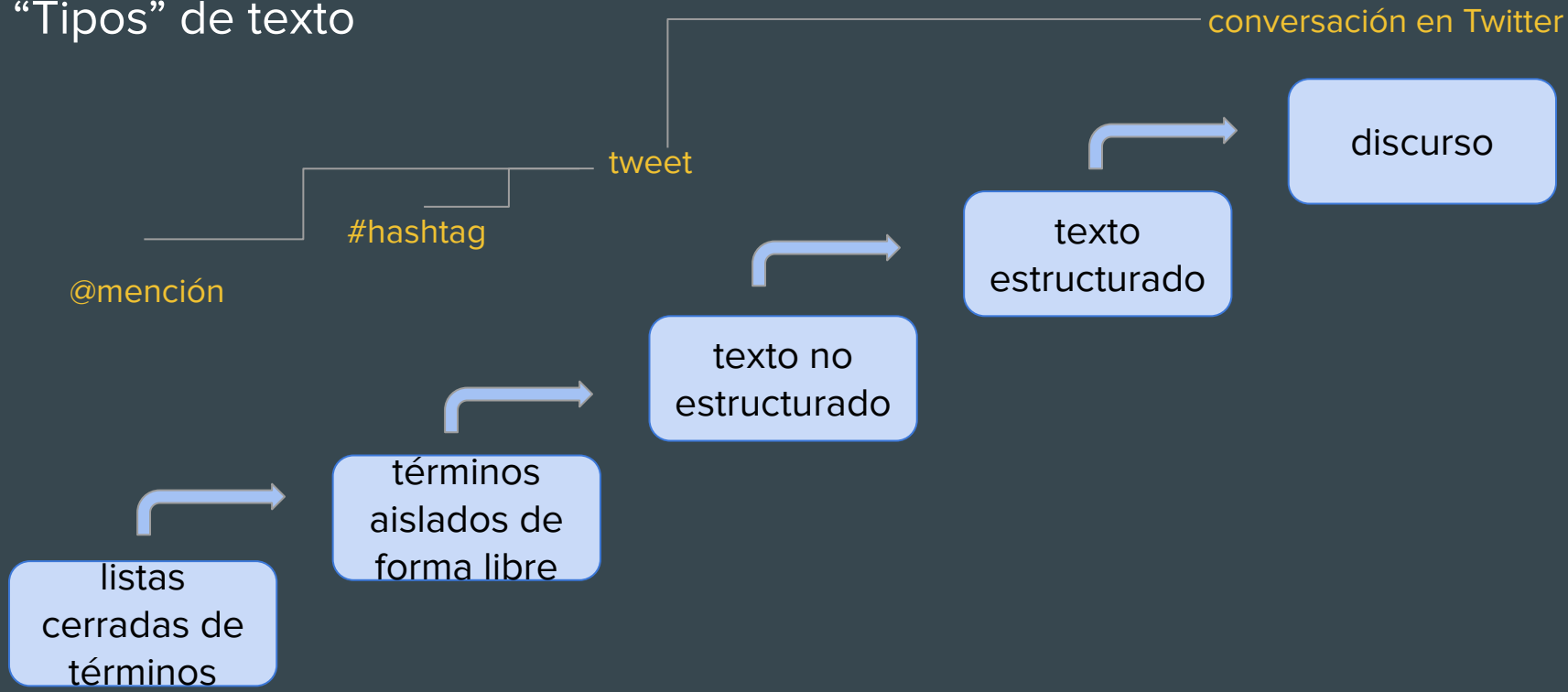
Procesado de texto

“Tipos” de texto



Procesado de texto

“Tipos” de texto



Procesado de texto

¿Qué podemos hacer con texto?

PoS tagging

analizador sintáctico

tokenization

análisis de patrones y estructuras

lemmatization

clasificación

separación y procesado de textos

identificación de topics

clustering

identificación del idioma

relaciones conceptuales

análisis de términos y sus relaciones

TF-IDF

pregunta-respuesta

ontologías

significado

conteo de palabras

NER

análisis de sentimiento

índice inverso

traducción automática

Algunos retos

Algunos retos

Entre lo que *pienso*,
lo que *quiero decir*,
lo que *creo decir*,
lo que *digo*,
lo que *quieres oír*,
lo que *oyes*,
lo que *crees entender*,
lo que *quieres entender*,
lo que *entiendes*

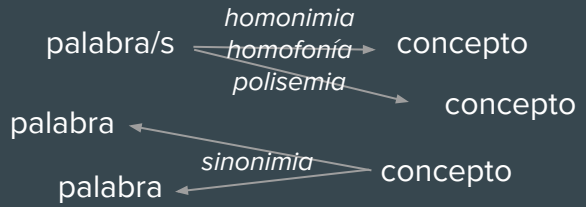
Existen 9 posibilidades de no entenderse

Algunos retos

Regla de la comunicación del 7-38-55 % de Mehrabian

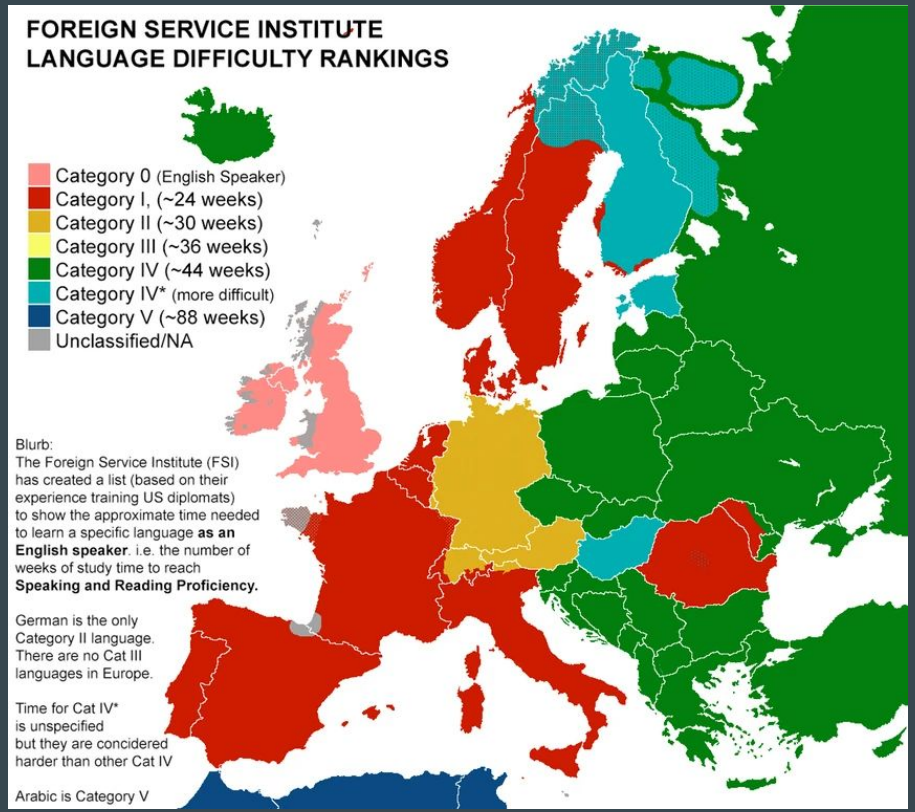
Algunos retos

Mapear palabras y conceptos

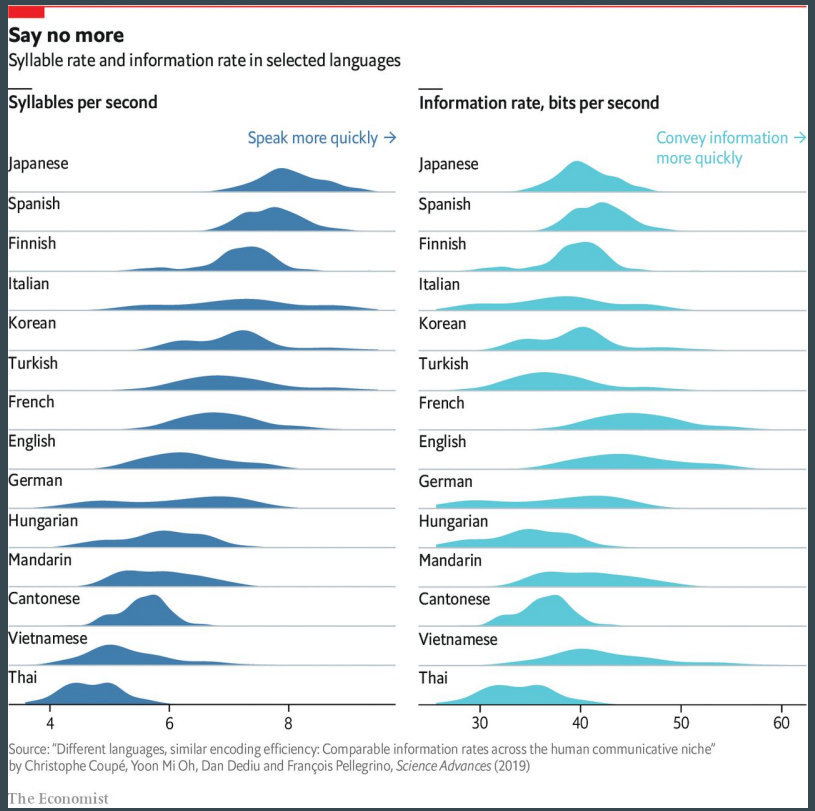


- Homonimia: misma forma, distinto significado, distinta etimología
  - Homografía: misma escritura
    - *Nada* (nadar) y *nada*, *vela* (velar) y *vela*, *bota* (calzado) y *bota* (vino), ...
  - Homofonía: misma pronunciación
    - *Basta* y *vasta*, *sabia* y *savia*, *botar* y *votar*, ...
- Polisemia: misma palabra (misma etimología), distintos significados
  - *Cabo* (escalafón militar) y *cabo* (accidente geográfico), *gato* (animal) y *gato* (herramienta), ...
- Sinonimia: relación de igualdad entre el significado de +2 palabras / enunciados
  - *Barato* y *económico*, *demente* y *loco*, *idioma* y *lengua*, ...

Algunos retos



Algunos retos



Algunos retos

- Typos
  - Ejemplo de respuestas en un formulario (¿dónde reside usted?): *lllescas*, *illescas*, *ILLESCAS*, *ILL*
- Ground truth
  - ¿Con qué entrenamos? ¿Con qué validamos?
- Tropos
  - Ironía, Sarcasmo, Metáfora, Alegoría, Oxímoron, ...
- Contexto
  - P: ¿Qué temperatura hace en Roma?
  - R: 27 °C
  - P: ¿Y en Lima?
  - R: ...
- Idioma
- Puede llegar cualquier texto (sobre todo en asistentes virtuales)

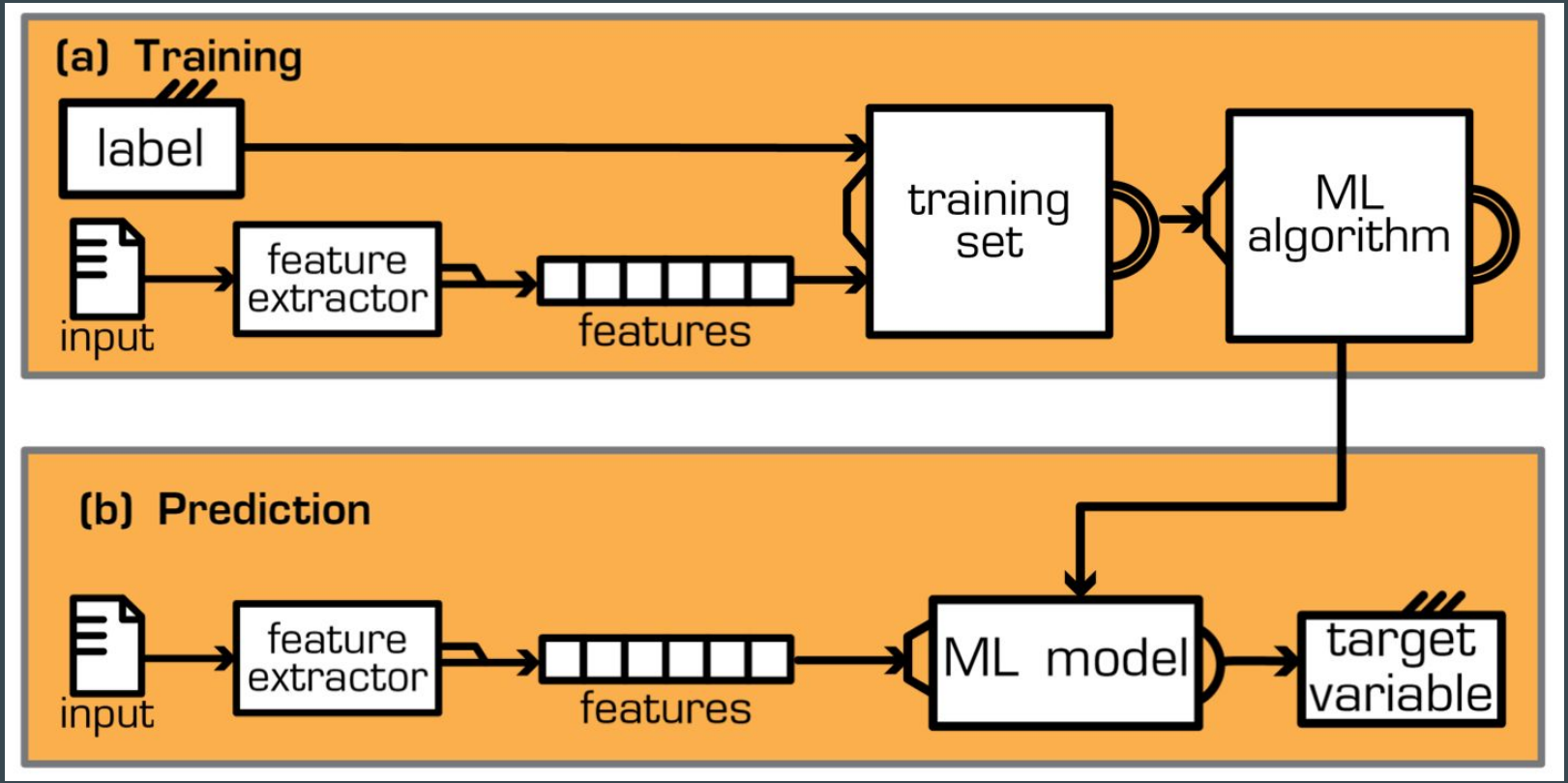
NLP vs Computer Vision

NLP vs CV

- En general, los modelos de CV mejor performance
- Modelos pre-entrenados
- Debate:
  - Retos en modelos de Visión Artificial
  - Retos en modelos de NLP

Pipeline de NLP

Pipeline


(a) Training
label
input
feature extractor
features
training set
ML algorithm
(b) Prediction
input
feature extractor
features
ML model
target variable

¡A la práctica!