

Modelos de Inteligencia Artificial

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



UNIDAD 3.

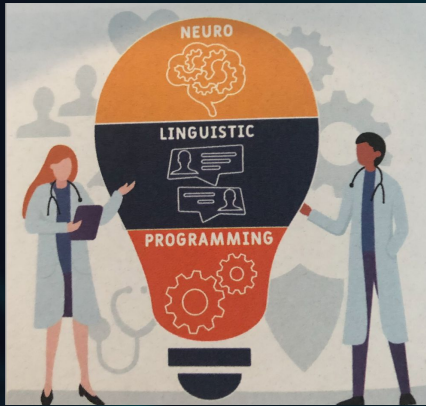
Procesamiento del Lenguaje Natural y sus aplicaciones. Potencial y limitaciones.

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

1. Procesamiento del Lenguaje Natural

Procesamiento del Lenguaje Natural




Estudia las relaciones del lenguaje entre los seres humanos y las máquinas.

Test de Georgetown (1954) → traducción automática de unas 60 oraciones del ruso al inglés.

¿Cómo se comunican las personas? → ***Lingüística***

Noam Chomsky

La teoría lingüística de Chomsky, conocida como innatismo, plantea el hecho de que, el lenguaje es producto de una facultad innata que posee la mente humana, que lo produce a través de estructuras predefinidas.



Objeto de estudio de la teoría lingüística de Chomsky

La teoría lingüística de Chomsky tiene como objetivo construir una Gramática Universal (GU) que constituye una teoría general de la estructura lingüística, independientemente de las lenguas particulares.



Más acerca de Noam Chomsky

Corpus

Conjunto de palabras (un texto) de una lengua que se emplea para formar un diccionario, pero no en el sentido de documento donde se explica o se definen las palabras, sino como concepto de conjunto de palabras de una lengua.

El corpus es la llave en la que se basa el aprendizaje máquina para el procesamiento del lenguaje.

Corpus del inglés → **Gutenberg**

<https://github.com/pgcorpus/gutenberg/>

Corpus de español

<https://github.com/roquegv/spanishNLPMoelCorpus>

Procesamiento del Lenguaje Natural

Será necesario disponer de un set de datos (corpus) específico para cada lengua → modelos específicos para cada idioma.

Necesario estudiar:

- La parte de la máquina
- La parte del ser humano

Según D. Jurafsky en su texto <<Speech and language processing>>:

- ❖ Estudio de las palabras y expresiones regulares del mismo,
- ❖ Relaciones generadas por:
 - Morfología
 - Sintaxis
 - Semántica
 - Pragmática

Papel del lingüista

En un proyecto de IA aporta el conocimiento y el enfoque para llevar a cabo exitosamente la programación de las complejas estructuras.

- Etiquetado → para clasificar a una palabra, grafía, sintagma o estructura determinada.

Ejemplo: Gato es la combinación de las grafías <<g>> <<a>> <<t>> <<o>> que a su vez desde un punto de vista semántico representa el concepto de un animal y cuya vocal <<o>> final denota el género de la palabra, que hace referencia al sexo del animal.

Conceptos

- **Gramática:** incluye la morfología, la sintaxis, y para algunos autores la fonología.
- **Sintaxis:** estudia las reglas y principios que gobiernan la combinación de los constituyentes sintácticos.
 - Analiza la formación de los sintagmas y las oraciones gramaticales.
 - Se conocen las formas en que se combinan las palabras, así como las relaciones sintagmáticas y paradigmáticas existentes entre ellas.
 - Como los sintagmas pueden unirse, para formar un grupo sintagmático, la sintaxis también establece la manera en la que llevar a cabo del proceso de Unificación.
 - En el etiquetado sintáctico se adjudica un POS (Part of Speech)

Ejemplo: para la oración <<el niño llora>>

El → determinante

Niño → Nombre

El niño → grupo sintagmático nominal

Llora → verbo (y grupo sintagmático predicativo)

- Dentro de una etiqueta pueden existir subcategorías:

Ejemplo: Niño → es un sustantivo (nombre) de tipo <<común>>

El → determinante de tipo artículo



Conceptos

- **Morfología:** según Jurafsky, estudia las reglas que rigen la flexión, la composición y la derivación de las palabras.
 - Durante el etiquetado morfológico se determina la **forma**, **clase** o **categoría gramatical** de una palabra.
 - El **género** de las palabras (masculino y femenino)
 - El **número** de los nombres (singular o plural)
 - La **persona de las formas conjugadas** (primera, segunda o tercera)
- **Semántica:** siguiendo el texto, estudia el significado de las expresiones lingüísticas, es decir, las realidades que representan las grafías.

Conceptos

- **Pragmática:** se centra en el análisis de la relación del lenguaje con los usuarios y las circunstancias de la comunicación o contexto.
 - El contexto debe entenderse como situación, ya que puede incluir cualquier aspecto extralingüístico: *la situación comunicativa, un conocimiento popular compartido por los habitantes, relaciones personales y otros muchos.*



Etiquetador para lengua española