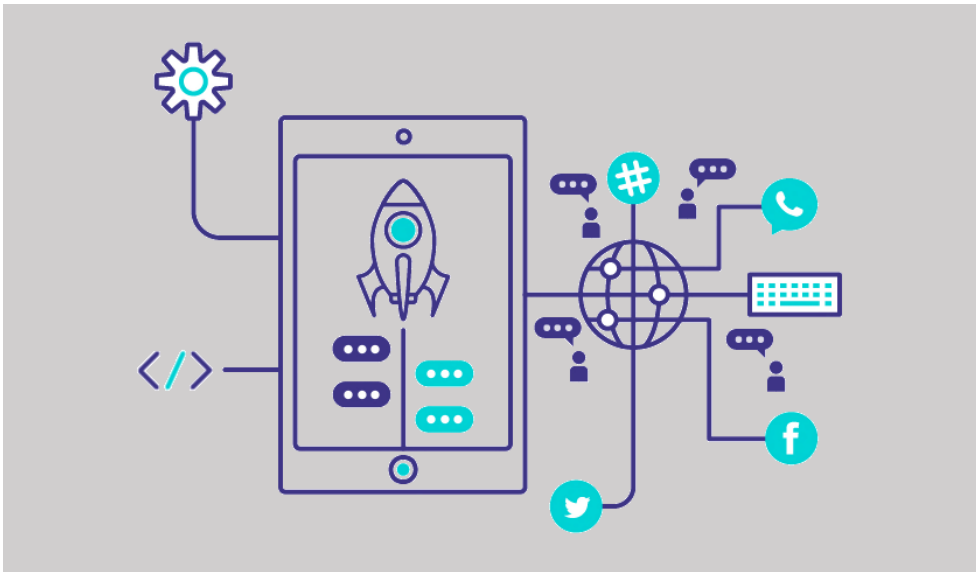


Inteligencia Artificial

Unidad 4: Procesamiento de Lenguaje Natural

TEMA 4: Algoritmos de IA Moderna-II

Módulo 2: Análisis del Lenguaje Natural

Profesora

Ing. Patricia Reyes Silva

Unidad 4

Procesamiento de Lenguaje Natural

TEMA 4: Algoritmos de IA Moderna-II

Sesión 24

MÓDULO 2:
Análisis del Lenguaje Natural

Contenido

1. Diferencias entre el Lenguaje Natural y el Lenguaje Formal
2. Algoritmo para la creación del Lenguaje Formal
3. La Teoría de los Lenguajes Formales
4. Componentes del NLP
5. Modelos para el procesamiento del Lenguaje Natural

Preguntas

1. Diferencias entre el lenguaje natural y el lenguaje formal

- Existen muchos tipo de lenguajes: hablado, escrito, gráfico, señas, sonidos, etc.

- En algunas ciencias, especialmente en **lógica** y **ciencias de la computación**, hay diferentes enfoques para el lenguaje.

- La **lógica simbólica**, estudia las reglas del correcto pensar, mismas que nos ayudan a expresar y ordenar nuestros pensamientos.

- Desde el punto de vista de la **lógica simbólica**, nos interesa diferenciar entre **lenguaje natural** y **lenguaje formal**.

1. Diferencias entre el lenguaje natural y el lenguaje formal

Lenguaje Natural

- No siempre cumple las reglas gramaticales y ortográficas
- Las oraciones suelen ser simples y cortas.
- Se dirigen al receptor de tú.
- Utiliza un vocabulario más bien pobre, repetitivo y reiterativo.
- Uso de muletillas, jergas, modismos o vulgarismos.
- La pronunciación no siempre es la correcta.
- A veces se omiten palabras de tal forma que la comunicación sea más rápida.
- Es un registro en el que abundan expresiones de carácter coloquial y con rasgos expresivos tales como juegos de palabras o frases hechas

Lenguaje Formal

Posee una gramática y ortografía correctas.

Las oraciones suelen ser largas y complejas.

- Utiliza un vocabulario rico y variado.
- Suelen dirigirse al receptor de usted.
- Utiliza sinónimos o pronombres para evitar redundancias.
- Correcta pronunciación.
- Evita expresiones como jergas, modismos, vulgarismos o muletillas.
- No existen omisiones.
- La información se presenta de forma estructurada y coherente.
- No acepta diminutivos o cualquier otro tipo de expresiones de carácter coloquial.

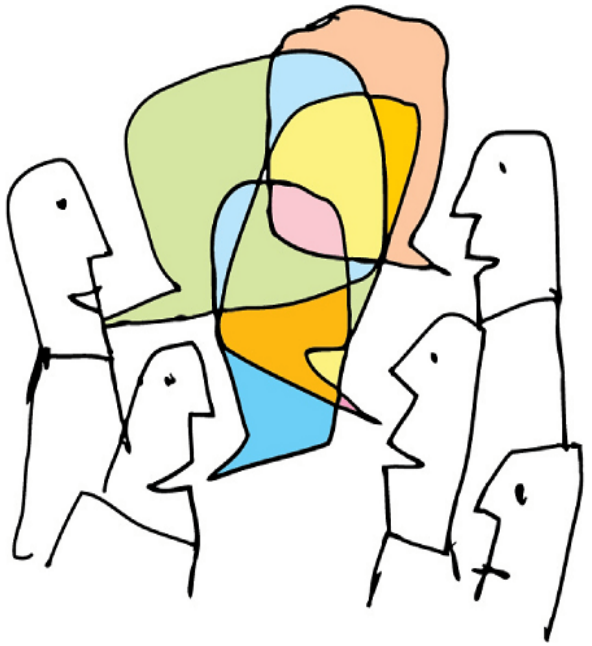
1. Diferencias entre el lenguaje natural y el lenguaje formal

- Como podemos ver, el lenguaje natural y formal son opuestos entre sí.

- A pesar de ello, el lenguaje natural y formal están estrechamente relacionados entre sí y tienen una serie de principios comunes. Estos son:

1. Se complementan entre sí (en mayor medida, lo natural complementa lo formal). Por lo tanto, hablar le permite crear la visión inicial de un proceso o un objeto.

2. Tanto el lenguaje natural como el formal inicialmente se basan en los mismos algoritmos. Esto es, comparten **la unidad básica del habla es un símbolo**, es decir, **una letra**. Del conjunto de caracteres se formaron palabras y frases.



2. Algoritmo para la creación del Lenguaje Formal

- **La teoría de los lenguajes formales** es una disciplina científica que en informática y lógica, estudia los objetos de un lenguaje formal.
- La mayoría de las construcciones en las que se usa un lenguaje formal, como regla, se construyen mediante el siguiente algoritmo:

1. En primer lugar, se selecciona un conjunto de símbolos necesarios (signos y letras) para el trabajo, es decir, el alfabeto.

2. A continuación, se especifican las reglas y principios para la formación de estructuras.

   Esto es necesario para construir la secuencia lógica correcta de caracteres que, en el futuro, crearán un componente de texto significativo.

   Aquí el factor de formación de palabras es importante.

3. Las palabras completas se obtienen gradualmente a partir de palabras individuales.

a: a	m: eme
b: be	n: ene
c: ce	ñ: eñe
ch: che	o: o
d: de	p: pe
e: e	q: cu
f: efe	r: erre
g: ge	s: ese
h: hache	t: te
i: i	u: u
j: jota	v: uve
k: ka	x: equis
l: ele	y: i griega
ll: elle	z: zeta

2. Algoritmo para la creación del Lenguaje Formal

- Existen dos grandes disciplinas matemáticas ocupadas de poner orden en el universo de los lenguajes formales:

❑ **La Teoría de los Lenguajes Formales (Sensu Strictu)**

Estudia los lenguajes prestando atención únicamente a sus propiedades estructurales, definiendo clases de complejidad estructural y estableciendo relaciones entre las diferentes clases.

❑ **La Teoría de la Complejidad Computacional**

Estudia los lenguajes prestando atención a los recursos que utilizaría un dispositivo mecánico para completar un procedimiento de decisión, definiendo así diferentes clases de complejidad computacional y las relaciones que existen entre ellas.

Pero, ¿Quiénes ponen orden en el universo de los lenguajes formales?

3. La Teoría de los Lenguajes Formales

Entendiendo los Lenguajes Formales

Los **lenguajes formales** los podemos representar como un conjunto, finito o infinito, de cadenas definidas sobre un alfabeto finito.

Ejemplo lenguaje: L_1 = {ab, aabb, aaabbb}
L_2 = {001, 011, 111}

a. Siendo el **Alfabeto** un conjunto finito de símbolos:
Σ_1 = {a, b, c}.

b. Donde cadena o palabra es una serie arbitraria de símbolos unidos.
Ejemplo **palabra**: aaabbbccc

Pero, no todos los lenguajes son de interés

- No todos los lenguajes que pueblan este universo son del interés de la teoría de los lenguajes formales.
- Sólo son interesantes aquellos lenguajes en los que se observa que se sigue alguna pauta regular en la construcción de las cadenas.

Desde este punto de vista, el lenguaje L_3 es digno de estudio, pero no L_4

L_3 = {abc, aabbccc, aaabbbccc, ...}

L_4 = {a, cab, bdac, ...}

3. La Teoría de los Lenguajes Formales

Entendiendo los Lenguajes Formales: Clausura de Kleene

Existe una denominación llamada ***clausura transitiva*** o, también, ***clausura de Kleene*** de un alfabeto **Σ**, escrito **Σ***, que es el conjunto de todas las cadenas sobre **Σ**.

Por tanto: {a}* = {ε, a, aa, aaa, ...}

{a, b}* = {ε, a, b, aa, bb, ab, ba, aaa, ...}

Propiedades de la clausura transitiva o de Kleene:

$\boldsymbol{L_1}$ = {x ϵ {a, b}* | |x| <= 2}

$\boldsymbol{L_2}$ = {xy | x ϵ {a, aa} e y ϵ {b, bb}}

$\boldsymbol{L_3}$ = $\{a_n\ b_n \mid n >= 1\}$

$\boldsymbol{L_4}$ = $\{(ab)_n \mid n >= 1\}$

3. La Teoría de los Lenguajes Formales

Entendiendo los Lenguajes Formales: Expresiones Regulares

- Una expresión regular es una fórmula **r** cuya denotación es un lenguaje **L(r)** definido sobre un alfabeto **Σ**.
- Hay dos tipos de expresiones regulares:
  - ❖ Las expresiones regulares atómicas.
  - ❖ Las expresiones regulares compuestas.

Expresiones regulares atómicas

Sintaxis y denotación

1. Cualquier literal **a**, tal que **a ϵ Σ** es una expresión regular cuya denotación es **L(a)** = **{a}**
2. El símbolo especial **ε** es una expresión regular cuya denotación es **L(ε)** = **{ε}**
3. El símbolo especial **φ** es una expresión regular cuya denotación es **L(φ)** = **{ }**

Expresiones regulares compuestas

Sintaxis y denotación

1. $(r_1\ r_2)$ es una expresión regular cuya denotación es:
   L $(\boldsymbol{r_1\ r_2})$ = $L(r_1)L(r_2)$
2. $(r_1 + r_2)$ es una expresión regular cuya denotación es:
   L$(\boldsymbol{r_1 + r_2})$ = $L(r_1)$ U $L(r_2)$
3. (r)* es una expresión regular cuya denotación es:
   L((r)*) = (L(r))*

3. La Teoría de los Lenguajes Formales

Entendiendo los Lenguajes Formales: Expresiones Regulares

Ejemplos :

A= aa* denotación A={a, aa, aaa, …}

B= (a + b)* denotación B={a, b}*

C= a* + b* denotación $\{a_n\ b_n \mid n >= 0\}$

D= a*b* denotación $\{a_n\ b_m \mid n, m >= 0\}$

4. Componentes del NLP

4. Componentes del NLP

1. Análisis morfológico o léxico

- La **MORFOLOGÍA** es el estudio de las palabras y sus partes.
- Su unidad más importante es el **morfema**, que se define como la **"unidad mínima de significado"**.
- Por ejemplo, consideremos una palabra como: **“ungentlemanly”** o **"poco caballeroso"** cuya estructura se preparó con prefijo, sufijo y palabra raíz como se muestra a continuación:

Análisis Morfológico

4. Componentes del NLP

1. Análisis morfológico o léxico

El análisis morfológico ayuda a comprender la estructura y el significado de la palabra durante el procesamiento del texto.

CAMPOS BASICOS DE LA MORFOLOGIA

1. **La inflexión**

- Es el proceso de cambiar la forma de una palabra para que exprese información como número, persona, caso, género, tiempo, estado de ánimo y aspecto, pero la categoría sintáctica de la palabra permanece sin cambios como:

car / coche / car**s** / coche**s**
table / mesa / table**s** / mesa**s**.

2. **La derivación**

- Es el proceso de cambiar la categoría sintáctica de la palabra. Los morfemas derivacionales se utilizan para cambiar las categorías gramaticales de las palabras. Por ejemplo, analicemos los morfemas 'er' y 'ly'

bake (cocinar) / bak**er** (cocinero) → cambia la categoría gramatical de verbo a sustantivo
quick (rápido) /quick**ly** (rápidamente)→ cambia la categoría gramatical de adjetivo a adverbio

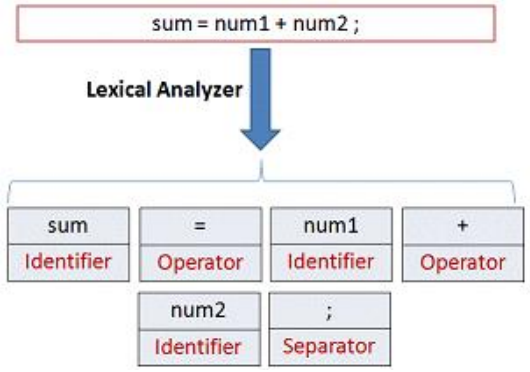
4. Componentes del NLP

1. Análisis morfológico o léxico

- El léxico de una lengua es su vocabulario que incluye sus palabras y expresiones.
- El análisis léxico consiste en dividir un texto en párrafos, palabras y oraciones.
- Se trata de comprender todo sobre las palabras distintas según su posición en el discurso, sus significados y su relación con otras palabras.

¿En qué consiste el análisis morfológico o léxico?

- Este análisis se realiza asociando, con cada palabra del léxico, información sobre los contextos en los que puede aparecer cada uno de los sentidos de la palabra.
- Por ejemplo, la declaración suma = num1 + num2 son procesados por Lexical Analyzer como se muestra en la imagen:



4. Componentes del NLP

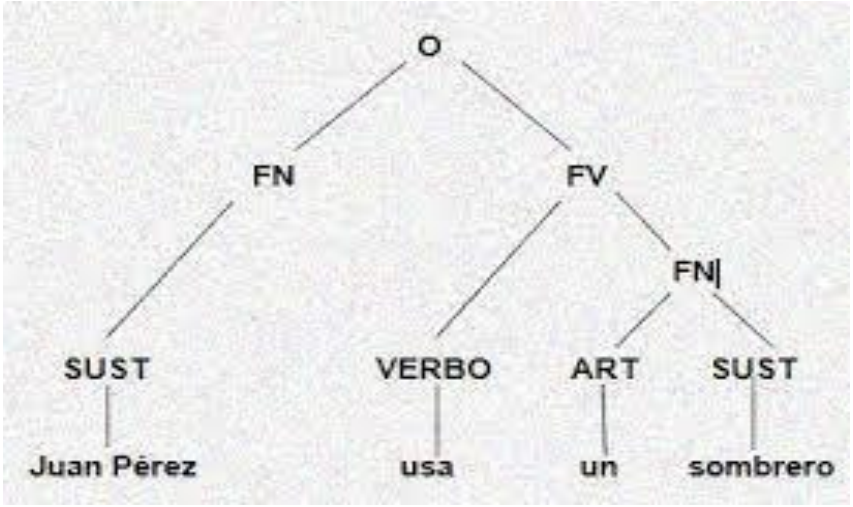
2. Análisis sintáctico

- El objetivo del análisis sintáctico es determinar si la cadena de texto en la entrada tiene la estructura correcta de oración / frase o no en el lenguaje (natural) dado.

- Si es correcta, entonces el resultado del análisis contiene una descripción de la estructura sintáctica de la frase, en la forma de un **árbol de derivación** o **árbol de análisis sintáctico** o **árbol de sintaxis**.

- Dichas formalizaciones (árboles) están destinadas a hacer que las computadoras "comprendan" las relaciones entre palabras (e indirectamente entre las personas, cosas y acciones correspondientes).

- Las representaciones sintácticas del lenguaje utilizan **gramáticas libres de contexto**, que muestran qué frases son partes de otras frases en lo que podría considerarse una forma libre de contexto.



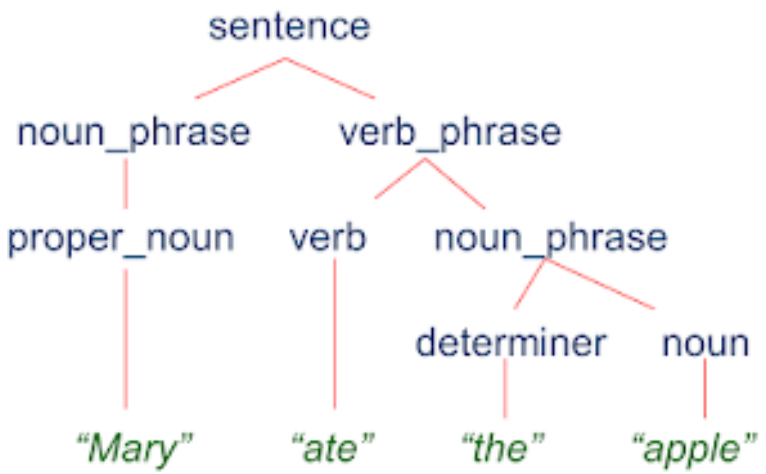
4. Componentes del NLP

2. Análisis sintáctico

USOS DEL ARBOL SINTACTICO

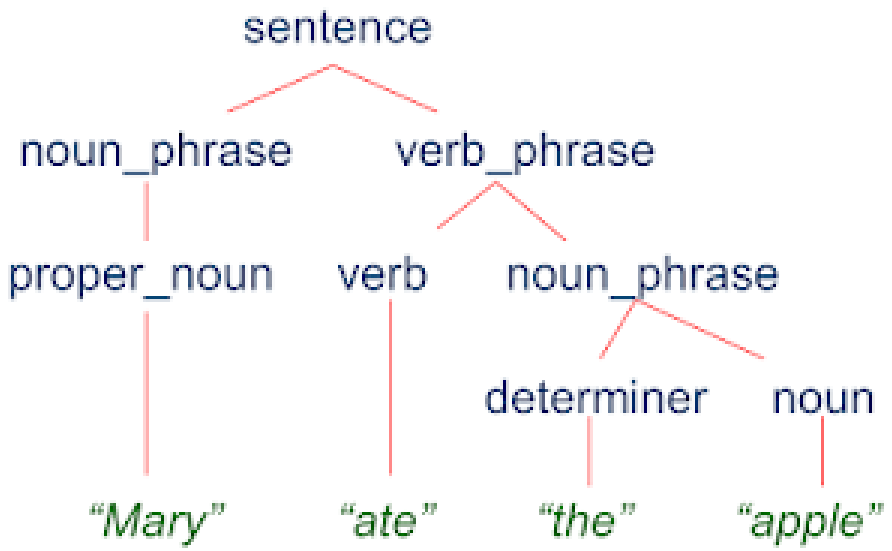
- Los **Parsing Tree** son directamente útiles en aplicaciones como la revisión gramatical en sistemas de procesamiento de texto, traducción automática, respuesta a preguntas y extracción de información.
- Una oración que no se puede analizar puede tener errores gramaticales (o al menos ser difícil de leer).
- Además, el análisis sintáctico es una etapa intermedia importante de representación para el análisis semántico.

- En el diagrama de la derecha se muestra el **árbol de análisis** de la oración "María se comió la manzana" utilizando el mecanismo de análisis de arriba hacia abajo.
- Aquí el árbol crece de arriba a abajo hasta que finalmente llegan a cada token o léxico y etiquetan cada palabra con su **etiqueta POS** (Parts Of Speech Tags) correspondiente utilizando un conjunto de reglas gramaticales.



4. Componentes del NLP

2. Análisis sintáctico



Para realizar un análisis sintáctico, necesitamos:

1. **Un analizador sintáctico:** un programa que toma como entrada una oración y produce el análisis.

2. **Una gramática:** un conjunto de reglas que puede usar el analizador. (Ver en el árbol como S -> NP VP,
   NP -> PPN (Sustantivo propio) ...)

3. **Un léxico:** un diccionario de palabras legales y sus partes del discurso (María, comió, la, manzana)

(su uso lo veremos en el ejemplo en Python con la librería NLTK)

4. Componentes del NLP

3. Análisis semántico

- Proporciona la interpretación de las oraciones, una vez eliminadas las ambigüedades morfosintácticas.
- Básicamente, es necesario completar un (1) análisis morfológico y (2) sintáctico antes de intentar resolver cualquier problema semántico.

¿Por qué es tan importante el análisis semántico para ofrecer contenido relevante?

Si se busca el término "jaguar", obtendrá resultados para:

Un depredador .. .

Un coche de lujo ...

Un equipo de fútbol ...

El análisis semántico ayuda a:

- Entregar el contenido que realmente se está buscando.
- Comprender tanto el contenido como la intención (o necesidad) del individuo es la clave para brindar una experiencia de usuario más valiosa y resonante.

4. Componentes del NLP

4. Análisis pragmático

- Incorpora el análisis del contexto de uso a la interpretación final.
- Aquí se incluye el tratamiento del lenguaje figurado (metáfora e ironía) como el conocimiento del mundo específico necesario para entender un texto especializado.

¿ Cómo ayuda el análisis pragmático a interpretar correctamente la oración?

- Pues teniendo en cuenta los siguientes factores:

1. **Utiliza el contexto del enunciado.**

   Dónde, por quién, a quién, por qué, cuándo se dijo
   Intenciones: informar, pedir, prometer , criticar ...

2. **Manejo de pronombres.**

   “María come manzanas. A ella le gustan. "
   Ella = "María", ellos = "manzanas".

3. **Manejo de la ambigüedad.**
   Ambigüedad pragmática: “llegas tarde”: ¿Cuál es la intención del hablante: informar o criticar?

5. Modelos para el procesamiento del Lenguaje Natural

- Tratar computacionalmente una lengua implica un proceso de **modelización matemática**.

- Los **lingüistas computacionales** se encargan de la tarea de **"preparar"** el modelo lingüístico para que los ingenieros informáticos lo implementen en un código eficiente y funcional.

- Existen dos aproximaciones generales al problema de la modelización lingüística:

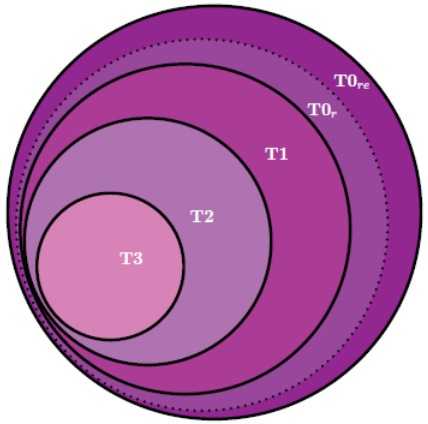
    1. **Modelos Lógicos:** gramáticas
    2. **Modelos Probabilísticos del Lenguaje Natural:** basados en datos (corpus)

5. Modelos para el procesamiento del Lenguaje Natural

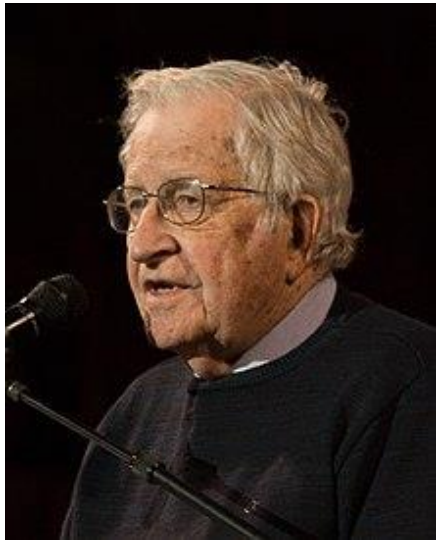
1. Modelos Lógicos: Gramáticas

- Estos modelos lógicos **pretenden reflejar la estructura lógica del lenguaje**.
- Estos modelos surgen a partir de las teorías de **N. Chomsky (Complejidad Estructural o Jerarquías de Chomsky)** en los años 50.
- Uno de los principales hallazgos de **Chomsky** fue la demostración de que podemos construir modelos matemáticos cuyas propiedades son un reflejo directo del grado de complejidad estructural de los lenguajes que se ajustan a dichos modelos.

Complejidad Estructural o Jerarquías de Chomsky



T3 = Lenguajes regulares.
T2 = Lenguajes independientes del contexto.
T1 = Lenguajes sensibles al contexto.
T0r = Lenguajes recursivos.
T0re = Lenguajes recursivamente enumerables.



Noam Chomsky lingüista, filósofo, politólogo y activista estadounidense de origen judío.

5. Modelos para el procesamiento del Lenguaje Natural

1. Modelos Lógicos: Gramáticas

A partir de este modelo lógico:

- Los lingüistas escriben **reglas de reconocimiento de patrones estructurales**, empleando un formalismo gramatical concreto.
- Luego, estas reglas, combinadas con la información almacenada en diccionarios computacionales, definen los patrones que hay que reconocer para resolver la tarea a resolver (buscar información, traducir, etc.).

Para Chomsky existen **reglas gramaticales universales** y otras muchas específicas de cada lengua, que permiten que los elementos que forman una oración puedan ordenarse de varias maneras.

Ejemplo: ‘Lapadula metió el gol' y 'Este gol ha sido metido por Lapadula’.

Según esto, Chomsky distingue dos **tipos de gramática**:

1. **TRANSFORMACIONAL:** Dispone unidades semánticas subyacentes, que mediante reglas las transforma en elementos de una oración, reconocible o interpretable.
2. **GENERATIVA:** Genera todas las oraciones aceptables.

5. Modelos para el procesamiento del Lenguaje Natural

1. Modelos Lógicos: Gramáticas

ROLES GRAMATICALES

La gramática, entendida como la técnica utilizada para formalizar un lenguaje, define:

- **FORMATOS:** Válidos para la combinación de símbolos de su alfabeto.
- **NORMAS:** Para generar el correspondiente lenguaje formal.
- **REGLAS:** Formadoras de las palabras componentes del lenguaje.

Ejemplo para estructurar frases en lenguaje español

Con estas reglas	Se generan las frases
< Frase > → < Sujeto > < Predicado > < Signo >	< Frase >→ **< Sujeto >** < Predicado > < Signo >
< Sujeto > → < Articulo > < Sustantivo >	→**< Articulo > < Sustantivo >** < Predicado > < Signo >
< Predicado> → < Verbo > < Objeto >	→< Articulo > < Sustantivo > **< Verbo > < Objeto >** < Signo>
< Objeto> → <Preposición><Sujeto>| λ |<Adverbio>	→< Articulo >< Sustantivo >< Verbo > **< Adverbio >**<Signo>
< Articulo> → | la | el | λ | los	→ **la Facultad enseña tecnología.**
< Verbo> → | enseña | ama | estudia	→< Articulo >< Sustantivo >< Verbo > **< Adverbio >**<Signo>
< Sustantivo> →| Facultad | λ | UPC | novio	→ **el novio ama mucho.**
< Adverbio> →| mucho | tarde | λ | medida|tecnología	
< Signo > → . | ; | ,	

5. Modelos para el procesamiento del Lenguaje Natural

1. Modelos Lógicos: Gramáticas

GRUPOS GRAMATICALES

Chomsky clasifica las gramaticas, en cuatro grupos				
Tipo	**Grupo**	**Gramatica**	**Lenguaje**	**Automata**
0	G_0	Sin restricción	L_0: **Recursivo enumerable** Nivel pragmático	Maquina de Turing MT
1	G_1	Dependiente del contexto	L_1: **Dependiente de contexto** Nivel Semántico	Autómata Linealmente Acotado ALA ó ALL
2	G_2	Independiente del contexto	L_2: **Independiente de contexto** Nivel Sintáctico	Autómata de Pila AP
3	G_3	Regular	L_3: **Regular** Nivel Léxico	Autómata Finito AF

La jerarquía implica que si L_1 contiene al conjunto L_0, en general, se define: $L_3 \subseteq L_2 \subseteq L_1 \subseteq L_0$

5. Modelos para el procesamiento del Lenguaje Natural

1. Modelos Lógicos: Gramáticas

GRAMATICA ESTRUCTURADA POR FRASES		
DEFINICION	**ELEMENTOS:**	**CONTENIDO:**
Es la **cuatrupla** (4-Tupla) $G=(\Sigma_T,\Sigma_N,S,P)$	• Σ_T:Alfabeto de símbolos terminales	Cada símbolo terminal conforma el formato no variable o final de la palabra
	• Σ_N:Alfabeto de símbolos no terminales	Cada símbolo no terminal es una **variable** que representa un estado intermedio de formación de la palabra, bajo la condición: $\Sigma=\Sigma_T\cup\Sigma_N$ y $\Sigma_T\cap\Sigma_N=\phi$
	• S: Axioma o símbolo inicial o de la gramática	Símbolo de inicio para generar cada palabra del lenguaje. Cumple con : $S \in \Sigma_N$
	• P: Reglas de Sustitución o Produccion	Conjunto finito de transformaciones para alcanzar el estado de palabra formada por solo símbolos terminales.Para esto: A partir del axioma S se detecta cada símbolo no terminal y se efectua la producción para transformarla en terminal.

Ejemplo: Si aplicamos los conceptos anteriores y usando paréntesis angulares < > para las estructuras sintácticas, **una gramática estructurada por frases**, podríamos definirla como:

$G = (\Sigma_T, \Sigma_N, S, P) =$ ({ el, la, los, Facultad, UPC, novios, crece, avanza, aman, para, con, sin, mucho, tarde, . , medida },
{<Frase><Sujeto>< Predicado><Signo><Articulo><Sustantivo><Verbo><Objeto>< Preposición >< Adverbio >},
<Frase>,P)

5. Modelos para el procesamiento del Lenguaje Natural

1. Modelos Lógicos: Gramáticas

GRAMATICA ESTRUCTURADA POR FRASES		
DEFINICION	**ELEMENTOS:**	**CONTENIDO:**
Es la **cuatrupla** (4-Tupla) $G=(\Sigma_T,\Sigma_N,S,P)$	• Σ_T:Alfabeto de símbolos terminales	Cada símbolo terminal conforma el formato no variable o final de la palabra
	• Σ_N:Alfabeto de símbolos no terminales	Cada símbolo no terminal es una **variable** que representa un estado intermedio de formación de la palabra, bajo la condición: $\Sigma=\Sigma_T\cup\Sigma_N$ y $\Sigma_T\cap\Sigma_N=\phi$
	• S: Axioma o símbolo inicial o de la gramática	Símbolo de inicio para generar cada palabra del lenguaje. Cumple con : $S \in \Sigma_N$
	• P: Reglas de Sustitución o Produccion	Conjunto finito de transformaciones para alcanzar el estado de palabra formada por solo símbolos terminales.Para esto: A partir del axioma S se detecta cada símbolo no terminal y se efectua la producción para transformarla en terminal.

Ejemplo: Si aplicamos los conceptos anteriores y usando paréntesis angulares < > para las estructuras sintácticas, **una gramática estructurada por frases**, podríamos definirla como:

$G = (\Sigma_T, \Sigma_N, S, P) =$ ({ el, la, los, Facultad, UPC, novios, crece, avanza, aman, para, con, sin, mucho, tarde, . , medida },
{<Frase><Sujeto>< Predicado><Signo><Articulo><Sustantivo><Verbo><Objeto>< Preposición >< Adverbio >},
<Frase>,P)

5. Modelos para el procesamiento del Lenguaje Natural

2. Modelos probabilísticos del lenguaje natural: basados en datos (CORPUS)

En estos modelos, la aproximación es a la inversa que en los modelos lógicos:

- Los lingüistas recogen colecciones de ejemplos y datos (corpus) y a partir de ellos se calculan las frecuencias de diferentes unidades lingüísticas (letras, palabras, oraciones) y su probabilidad de aparecer en un contexto determinado.
- Calculando esta probabilidad, se puede predecir cuál será la siguiente unidad en un contexto dado, sin necesidad de recurrir a reglas gramaticales explícitas.

- De esta manera, los algoritmos infieren las posibles respuestas a partir de los datos observados anteriormente en el corpus.

PREGUNTAS

Dudas y opiniones