Procesamiento del Lenguaje Natural

Procesamiento del Lenguaje Natural

Los agentes envían señales a otros agentes para…

- Informar
- Pedir ayuda
- Compartir conocimiento
- Preguntar
- Ordenar
- Comprometerse
- Etc.

Procesamiento del Lenguaje Natural

- La comunicación requiere ciertas convenciones que forman el lenguaje
- El lenguaje está formado por signos con significados acordados
- No es una habilidad propia de los humanos
- Los humanos parecen usan la gramática para producir infinitos mensajes estructurados

Procesamiento del Lenguaje Natural

Lenguajes formales

- Son un conjunto de cadenas
- Cada cadena es una concatenación de símbolos terminales (palabras)
- Ejemplos: lógica de predicados, Python
- **Tienen definiciones estrictas**
- Asocian un significado o semántica a cada cadena válida
- El significado de una cadena sólo está influido por su forma

Procesamiento del Lenguaje Natural

Lenguajes naturales

- Generado espontáneamente en un grupo de hablantes
- **No tiene definiciones estrictas**
- El significado específico y contextual de sus componentes intervienen en la validez o no de la frase
- El lenguaje natural no puede ser fácilmente caracterizado porque:
  - No existen reglas gramaticales claras (o no se respetan)
  - Hay ambigüedad
  - La relación entre símbolo y objetos no está formalmente definida
- Pragmática: el significado real de la cadena cuando es dicha en una situación determinada.
- La teoría de lenguajes formales es útil para el estudio de los lenguajes naturales

Procesamiento del Lenguaje Natural

Gramáticas

- Una gramática es un conjunto finito de reglas que especifican un lenguaje
- Define la sintaxis de las cadenas válidas
- Las cadenas están compuestas de subcadenas llamadas frases, las cuales pueden pertenecer a distintas categorías.
- Se usan tanto para interpretar el texto como para generarlo

Procesamiento del Lenguaje Natural

Categorías gramaticales

- Frases nominales (FN): Hacen referencia a objetos del mundo
- Frases verbales (FV): Indican Acciones
- A frases de cierta categoría, como las nominales y las verbales se puede combinar para crear frases del tipo "sentencia" (S).
- A las categorías (FV, FN, S, etc.) se las llama "símbolos no terminales"
- Las gramáticas definen los no terminales por medio de reglas de reescritura.
- Notación de la forma Backus-Naur (BNF): S→ FN FV (una sentencia puede consistir en una FN seguida por cualquier FV)

Procesamiento del Lenguaje Natural

Ejemplo del libro

Procesamiento del Lenguaje Natural

Ejemplo del libro

Sustantivo	→	**hedor** | **brisa** | **resplandor** | **nada** | **agente** | ***wumpus*** | **foso** | **oro** | **este** | …
Verbo	→	**es** | **ver** | **oler** | **disparar** | **sentir** | **heder** | **ir** | **tomar** | **llevar** | **matar** | **girar** | …
Adjetivo	→	**derecho** | **izquierdo** | **oriental** | **muerto** | **posterior** | **hediondo** | …
Adverbio	→	**aquí** | **allí** | **cerca** | **adelante** | **correctamente** | **a la izquierda** | **al este** | **al sur** | **atrás** | …
Pronombre	→	**mí** | **tú** | **yo** | **ello** | …
Nombre	→	**Juan** | **María** | **Alicante** | **Aristóteles** | …
Artículo	→	**el** | **un** | …
Preposición	→	**a** | **en** | **sobre** | …
Conjunción	→	**y** | **o** | **pero** | …
Dígito	→	**0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9**

Figura 22.3 El léxico para $\mathcal{E}_0$.

Procesamiento del Lenguaje Natural

S	$\rightarrow$	*FN FV*	Yo + siento una brisa
	|	*S Conjunción S*	Yo siento una brisa + y + yo huelo a *wumpus*
FN	$\rightarrow$	*Pronombre*	Yo
	|	*Nombre*	Juan
	|	*Sustantivo*	foso
	|	*Artículo* del *sustantivo*	el + *wumpus*
	|	*Dígito Dígito*	3 4
	|	*FN FP*	el *wumpus* + al este
	|	*FN OracionRel*	el *wumpus* + que es hediondo
FV	$\rightarrow$	*Verbo*	huele mal
	|	*FV FN*	siento + una brisa
	|	*FV Adjetivo*	es hediondo
	|	*FV FP*	gira + al este
	|	*FV Advervio*	ve + adelante
FP	$\rightarrow$	*Preposición FN*	a + el este
OracionRel	$\rightarrow$	**que** *FV*	que + es hediondo

Figura 22.4 La gramática para $\mathcal{E}_0$, con frases de ejemplo para cada regla.

Procesamiento del Lenguaje Natural

Sustantivo	→	**hedor** | **brisa** | **resplandor** | **nada** | **agente** | ***wumpus*** | **foso** | **oro** | **este** | …
Verbo	→	**es** | **ver** | **oler** | **disparar** | **sentir** | **heder** | **ir** | **tomar** | **llevar** | **matar** | **girar** | …
Adjetivo	→	**derecho** | **izquierdo** | **oriental** | **muerto** | **posterior** | **hediondo** | …
Adverbio	→	**aquí** | **allí** | **cerca** | **adelante** | **correctamente** | **a la izquierda** | **al este** | **al sur** | **atrás** | …
Pronombre	→	**mí** | **tú** | **yo** | **ello** | …
Nombre	→	**Juan** | **María** | **Alicante** | **Aristóteles** | …
Artículo	→	**el** | **un** | …
Preposición	→	**a** | **en** | **sobre** | …
Conjunción	→	**y** | **o** | **pero** | …
Dígito	→	**0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9**

Figura 22.3 El léxico para $\mathcal{E}_0$.

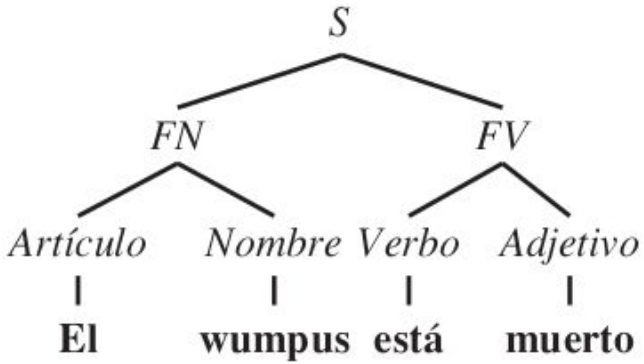
S	→	*FN FV*	Yo + siento una brisa
	|	*S Conjunción S*	Yo siento una brisa + y + yo huelo a *wumpus*
FN	→	*Pronombre*	Yo
	|	*Nombre*	Juan
	|	*Sustantivo*	foso
	|	*Artículo* del *sustantivo*	el + *wumpus*
	|	*Dígito Dígito*	3 4
	|	*FN FP*	el *wumpus* + al este
	|	*FN OracionRel*	el *wumpus* + que es hediondo
FV	→	*Verbo*	huele mal
	|	*FV FN*	siento + una brisa
	|	*FV Adjetivo*	es hediondo
	|	*FV FP*	gira + al este
	|	*FV Advervio*	ve + adelante
FP	→	*Preposición FN*	a + el este
OracionRel	→	**que** *FV*	que + es hediondo

Figura 22.4 La gramática para $\mathcal{E}_0$, con frases de ejemplo para cada regla.

Procesamiento del Lenguaje Natural

Análisis sintáctico (parsing)

- Descubrimiento de la estructura de las frases de acuerdo con las reglas de una gramática
- Búsqueda de un árbol de análisis válido cuyas hojas son las palabras de la cadena



Procesamiento del Lenguaje Natural

Gramáticas aumentadas

- Hasta ahora hemos visto gramáticas libres de contexto
- Las desventajas se hacen evidentes en la generación
  - No todas las frases sustantivas (por ejemplo) pueden aparecer en cualquier lugar. "Me comí una manzana" vs "Me comí una consulta de SQL"
  - Tampoco da lo mismo usar cualquier artículo. "La casa" vs "Los casa"

Procesamiento del Lenguaje Natural

Gramáticas aumentadas

- Hay más reglas que podríamos agregar. Por ejemplo:
  - Una gramática simple puede generar “Yo huelo un hedor”, pero también genera “Mi huelo un hedor”
  - “Yo” se emplea en el caso subjetivo y “mi” en el caso objetivo

$$
\begin{aligned}
S &\rightarrow FN_S\ FV \mid \ldots \\
FN_S &\rightarrow Pronombre_S \mid Nombre \mid Sustantivo \mid \ldots \\
FN_O &\rightarrow Pronombre_O \mid Nombre \mid Sustantivo \mid \ldots \\
FV &\rightarrow FV\ FN_O \mid \ldots \\
FP &\rightarrow Preposición\ FN_O \\
Pronombre_S &\rightarrow \textbf{Yo} \mid \textbf{tu} \mid \textbf{el} \mid \textbf{ella} \mid \textbf{ello} \mid \ldots \\
Pronombre_O &\rightarrow \textbf{mi} \mid \textbf{tu} \mid \textbf{su} \mid \ldots
\end{aligned}
$$

Procesamiento del Lenguaje Natural

Gramáticas aumentadas

- Con los pronombres por ejemplo, además sabemos que "Yo" es la 1ra persona del singular, "Ustedes" es la segunda del plural, etc.
- Una categoría como Pronombre que ha sido aumentada con características como el género y número es llamada subcategoría.
- Estas gramáticas pueden representar ese tipo de conocimiento para hacer una distinción más específica de qué tan probable es una oración.

Procesamiento del Lenguaje Natural

Gramáticas aumentadas

- Aumentar la cantidad de reglas para incluir a cada subcategoría haría crecer exponencialmente el tamaño de la gramática.
- En vez de introducir nuevas reglas. Las reglas aumentadas permiten parámetros para las categorías no terminales.

Procesamiento del Lenguaje Natural

Gramáticas aumentadas

- Aumentar la [illegible] ría haría crecer exponencial
- En vez de ir [illegible] ermiten parámetros

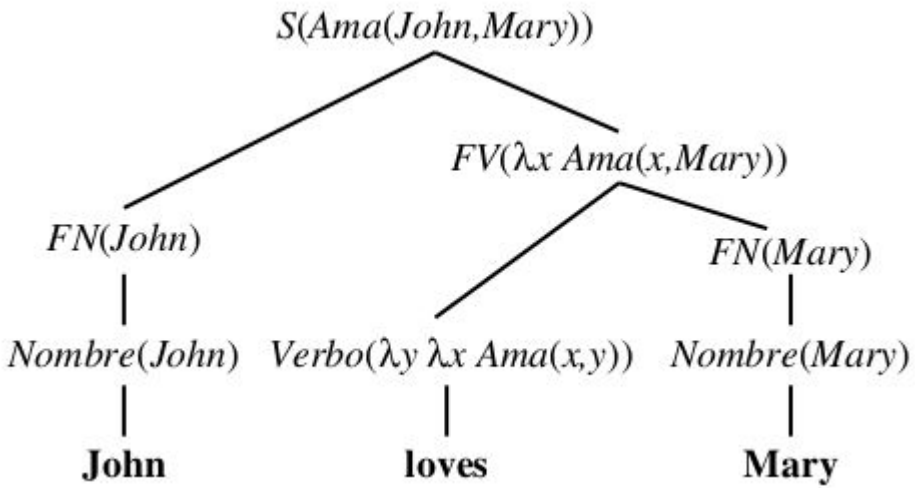
$$
\begin{aligned}
S &\rightarrow FN_S\ FV \mid \ldots \\
FN_S &\rightarrow Pronombre_S \mid Nombre \mid Sustantivo \mid \ldots \\
FN_O &\rightarrow Pronombre_O \mid Nombre \mid Sustantivo \mid \ldots \\
FV &\rightarrow FV\ FN_O \mid \ldots \\
FP &\rightarrow Preposición\ FN_O \\
Pronombre_S &\rightarrow \textbf{Yo} \mid \textbf{tu} \mid \textbf{el} \mid \textbf{ella} \mid \textbf{ello} \mid \ldots \\
Pronombre_O &\rightarrow \textbf{mi} \mid \textbf{tu} \mid \textbf{su} \mid \ldots
\end{aligned}
$$

$$
\begin{aligned}
S &\rightarrow FN(Subjetivo)\ FV \mid \ldots \\
FN(caso) &\rightarrow Pronombre(caso) \mid Nombre \mid Sustantivo \mid \ldots \\
FV &\rightarrow FV\ FN(Objetivo) \mid \ldots \\
FP &\rightarrow Preposición\ FN(Objetivo) \\
Pronombre(Subjetivo) &\rightarrow \textbf{Yo} \mid \textbf{tu} \mid \textbf{el} \mid \textbf{ella} \mid \textbf{ello} \mid \ldots \\
Pronombre(Objetivo) &\rightarrow \textbf{mi} \mid \textbf{tu} \mid \textbf{su} \mid \ldots
\end{aligned}
$$

Procesamiento del Lenguaje Natural

Interpretación semántica

- Se incorporan en la gramática equivalencias semánticas



Procesamiento del Lenguaje Natural

Modelos probabilísticos de lenguaje

- En lenguaje natural no se puede definir (como verdadero/falso) si una cadena es gramatical, pero podemos calcular qué tan probable es que una lo sea
- Un modelo de lenguaje es una distribución que describe la probabilidad de cualquier cadena de texto.
- ¿Probabilidad de qué? De ser gramaticalmente correcta, de ser una respuesta adecuada, de representar una situación dada, una traducción, etc..

Procesamiento del Lenguaje Natural

Modelos probabilísticos de lenguaje

- Bag-of-words
- Modelos n-gram
- Gramáticas

Procesamiento del Lenguaje Natural

Bag-of-words

- Permite clasificar textos
- Es un modelo generativo

$$\mathbf{P}(Class \mid w_{1:N}) = \alpha \, \mathbf{P}(Class) \prod_j \mathbf{P}(w_j \mid Class)$$

Procesamiento del Lenguaje Natural

N-gram word models

- Una de las limitaciones de bag-of-words es que hay palabras que pueden pertenecer a más de un dominio, por ejemplo, “unicornio”
- Si se toman conjuntos de n palabras adyacentes es más fácil determinar la clase o predecir la próxima palabra

$$P(w_j \mid w_{1:j-1}) = P(w_j \mid w_{j-n+1:j-1})$$

$$P(w_{1:N}) = \prod_{j=1}^{N} P(w_j \mid w_{j-n+1:j-1})$$

Procesamiento del Lenguaje Natural

Gramáticas

S	$\rightarrow$	$NP\ VP$	[0.90]	I + feel a breeze
	|	$S\ Conj\ S$	[0.10]	I feel a breeze + and + It stinks
NP	$\rightarrow$	$Pronoun$	[0.25]	I
	|	$Name$	[0.10]	Ali
	|	$Noun$	[0.10]	pits
	|	$Article\ Noun$	[0.25]	the + wumpus
	|	$Article\ Adjs\ Noun$	[0.05]	the + smelly dead + wumpus
	|	$Digit\ Digit$	[0.05]	3 4
	|	$NP\ PP$	[0.10]	the wumpus + in 1 3
	|	$NP\ RelClause$	[0.05]	the wumpus + that is smelly
	|	$NP\ Conj\ NP$	[0.05]	the wumpus + and + I
VP	$\rightarrow$	$Verb$	[0.40]	stinks
	|	$VP\ NP$	[0.35]	feel + a breeze

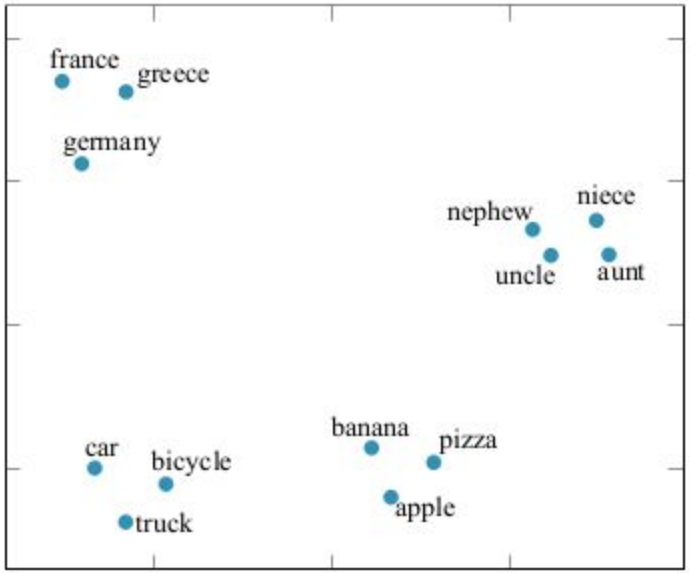
Procesamiento del Lenguaje Natural

Deep learning

- Dada la cantidad de texto disponible, tenía sentido considerar enfoques basados en aprendizaje automático orientado a datos.
- Algunas alternativas:
  - Word embeddings
  - Redes recurrentes
  - LSTM
  - Transformers

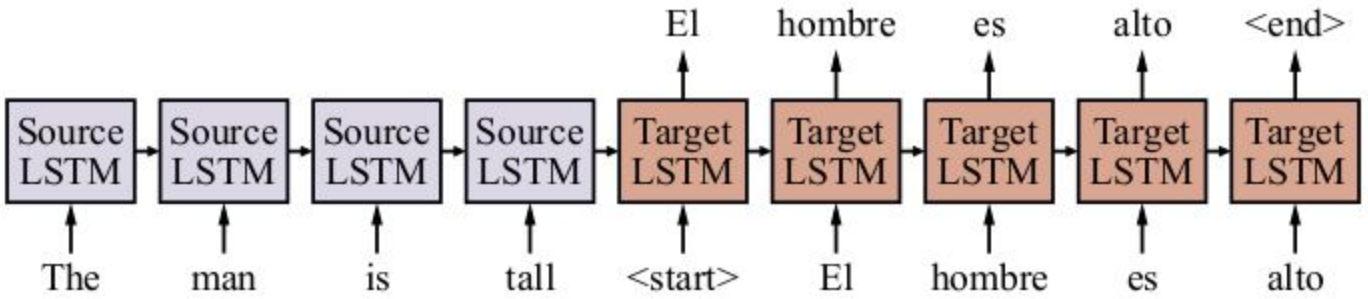
Procesamiento del Lenguaje Natural

Word embeddings



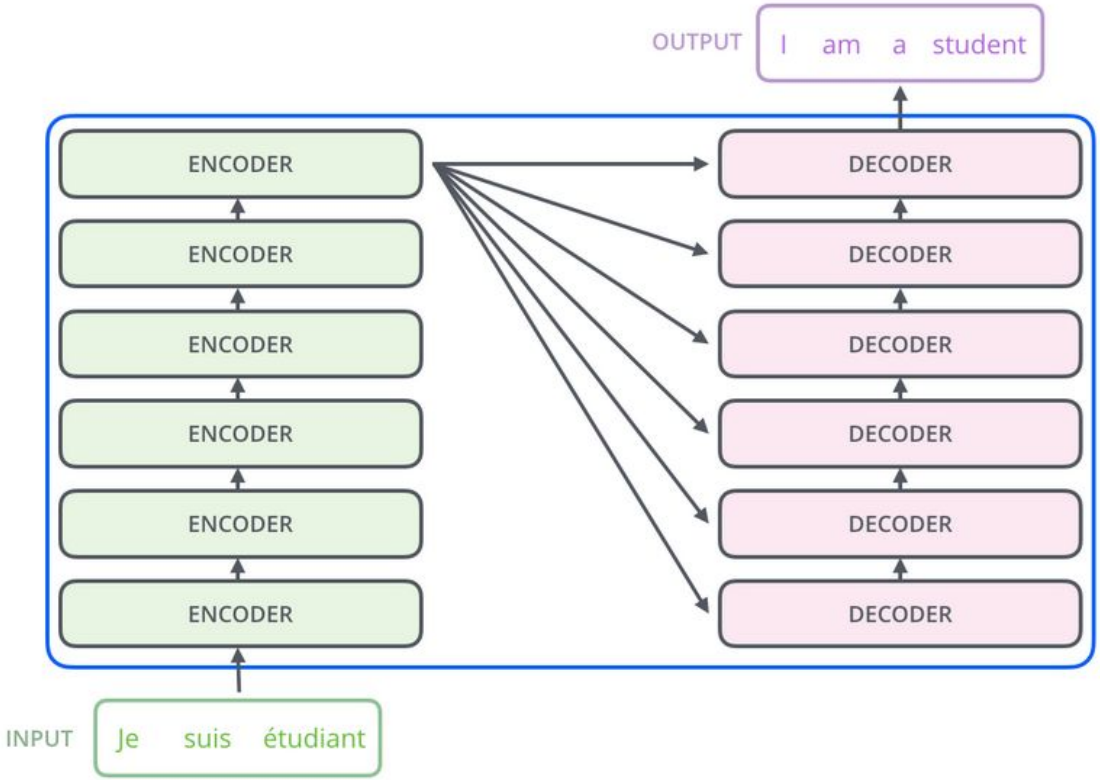
Procesamiento del Lenguaje Natural

LSTM - Sequence-to-sequence



Procesamiento del Lenguaje Natural

Transformers



Fin