Introducción al Procesamiento de Lenguaje Natural

Noviembre 2021 (Solución)

Consideraciones generales	
i)	La prueba es sin material escrito.
ii)	La duración es de 3 horas
iii)	Escriba nombre y CI al inicio
iv)	El total de puntos es 70

Ejercicio 1 [21 puntos]

a) Para cada afirmación diga si es Verdadera o Falsa. Justifique los casos de que su respuesta sea Falso.

i) La arquitectura Encoder-Decoder se utiliza para la traducción automática.

Verdadero. La red encoder codifica la oración en idioma origen, y la red decoder construye una oración en el idioma destino.

ii) WordNet es un algoritmo de desambiguación semántica de las palabras.

Falso. WordNet es una base de datos léxica (u ontología léxica) que contiene synsets, o sea conceptos o unidades únicas de sentido de las palabras.

iii) El Recall es una medida que puede ser usada para validar la similitud de dos documentos.

Falso. El ***recall*** es una medida que indica la cantidad de documentos recuperados relevantes sobre el total de todos los relevantes

iv) La lematización de una palabra ***x*** es el proceso que permite generar todas las posibles palabras derivadas de esa palabra ***x***.

Falso. La lematización es el proceso que consiste en llevar las palabras a una forma canónica (representativa); como por ejemplo para todas las ocurrencias de un mismo verbo, llevarlas a su infinitivo.

v) El algoritmo de parsing CKY funciona como un validador de oraciones sintácticamente correctas.

Verdadero. Es un algoritmo de análisis sintáctico.

vi) En la frase *La casa de papel* la palabra ***de*** tiene categoría adverbio.

Falso. La palabra de es una preposición.

vii) En la frase *Juan se sentó bajo el árbol* la palabra **bajo** tiene la categoría preposición.

Verdadero. Si bien la palabra bajo puede tener asociada varias categorías gramaticales según su uso, en este ejemplo cumple la función de preposición.

Ejercicio 2

a) ¿Qué es un constituyente en una gramática? Mencione al menos dos constituyentes de la gramática del español. ¿Qué nombre reciben los constituyentes en una gramática de dependencias?

Se le llama constituyente sintáctico a la palabra o conjunto de palabras que funcionan como una unidad en la estructura de una oración; como por ejemplo Grupo o Sintagma Nominal y Grupo o Sintagma Verbal.

En una gramática de dependencias sin embargo, no existe el concepto de constituyente, sino que se construye un grafo con etiquetas, que representan relaciones biléxicas entre las palabras, bajo el concepto de que una palabra "gobierna" a otra.

b) Considere una gramática con el siguiente conjunto de reglas de producción:

O → GN GV | GV

GV → V GN | V | V GN GP

GN → Det Nom | Nom | GN GP

GP → Prep GN

Nom → verano | primavera | caramelo | luna

Det → el | la

Prep → de | desde

V → miro | como

Sea la oración: *"Miro la luna de verano"*

i) Realice el análisis sintáctico aplicando el algoritmo de Earley

Chart[0]

γ → · O [0,0] Dummy
O → · GN GV [0,0] Predict
O → · GV [0,0] Predict
GN →· Det Nom [0,0] Predict

GN →· Nom [0,0] Predict
GN →· GN GP [0,0] Predict
GV →· V [0,0] Predict
GV →· V GN [0,0] Predict
GV →· V GN GP [0,0] Predict

Chart[1]
V → miro · [0,1] Scan
GV → V · [0,1] Complete
GV → V · GN [0,1] Complete
GV → V · GN GP [0,1] Complete
O → GV · [0,1] Complete
γ → O · [0,1] Complete
GN →· Det Nom [1,1] Predict
GN →· Nom [1,1] Predict
GN →· GN GP [1,1] Predict

Chart[2]
Det → la · [1,2] Scan
GN →Det · Nom [1,2] Complete

Chart[3]
Nom → luna · [2,3] Scan
GN → Det Nom · [1,3] Complete
GN → GN · GP [1,3] Complete
GP → · Prep GN [3,3] Predict
GV → V GN · [0,3] Complete
GV → V GN· GP [0,3] Complete
O → GV · [0,3] Complete
γ → O · [0,3] Complete

Chart[4]
Prep → de · [3,4] Scan
GP → Prep· GN [3,4] Complete
GN →· Det Nom [4,4]Predict
GN →· Nom [4,4] Predict
GN →· GN GP [4,4] Predict

Chart[5]
Nom → verano · [4,5] Scan
GN → Nom · [4,5] Complete
GN → GN · GP [4,5] Complete
GN → GN GP· [1,5] Complete
GP → Prep GN· [3,5] Complete
O → GV· [0,5] Complete
γ → O · [0,5] Complete
GV → V GN· [0,5] Complete
GV → V GN· GP [0,5]Complete
GP → · Prep GN [5,5]Predict

La oración entonces podemos afirmar que es generada por la gramática.

ii) Realice el análisis utilizando la notación bracket

[Miro [la luna]$_{GN}$ [de verano]$_{GP}$] $_{GV}$

Ejercicio 3

a) Sea la siguiente gramática con anotaciones semánticas:

o → gn gv	o.sem = gn.sem(gv.sem)
gn → det nom	gn.sem = det.sem(nom.sem)
gn → npropio	gn.sem = npropio.sem
gv → v	gv.sem = v.sem
nom → n	nom.sem = n.sem
nom → n adj	nom.sem = adj.sem(n.sem)
det → un	det.sem = λP. λQ. ∃x P(x) ∧ Q(x)
det → todo	det.sem = λP. λQ. ∀x P(x) → Q(x)
n → libro	n.sem = λx.libro(x)
n → pájaro	n.sem = λx.pájaro(x)
npropio → Pedro	npropio.sem = λP.P(pedro)
v → habla	v.sem = λx. habla(x)
adj → verde	adj.sem = λP.λx. verde(x) ∧ P(x)

Utilizando las reglas anteriores realice una derivación para calcular la representación semántica de la oración:

Todo pájaro verde habla

o.sem = gn.sem(gv.sem) *[SUSTITUCIÓN]*

= det.sem(nom.sem)(v.sem) *[SUSTITUCIÓN]*

= (λP . λQ . ∀x . P(x) → Q(x)) (adj.sem(n.sem)) (λx . habla(x)) *[SUSTITUCIÓN]*

= (λP . λQ . ∀x . P(x) → Q(x)) ((λP . λx . verde(x) ∧ P(x))(λx . pájaro(x))) (λx . habla(x)) *[SUSTITUCIÓN]*

= (λP . λQ . ∀x . P(x) → Q(x)) ((λP . λx . verde(x) ∧ P(x))(λy . pájaro(y))) (λx . habla(x)) *[CAMBIO DE VARIABLE]*

= (λP . λQ . ∀x . P(x) → Q(x)) (λx . verde(x) ∧ (λy . pájaro(y))(x)) (λx . habla(x)) *[APLICACIÓN FUNCIONAL]*

= (λP . λQ . ∀x . P(x) → Q(x)) (λx . verde(x) ∧ pájaro(x)) (λx . habla(x)) *[APLICACIÓN FUNCIONAL]*

= (λP . λQ . ∀x . P(x) → Q(x)) (λy . verde(y) ∧ pájaro(y)) (λx . habla(x)) *[CAMBIO DE VARIABLE]*

= (λQ . ∀x . (λy . verde(y) ∧ pájaro(y))(x) → Q(x)) (λx . habla(x)) *[APLICACIÓN FUNCIONAL]*

= (λQ . ∀x . verde(x) ∧ pájaro(x) → Q(x)) (λx . habla(x)) *[APLICACIÓN FUNCIONAL]*

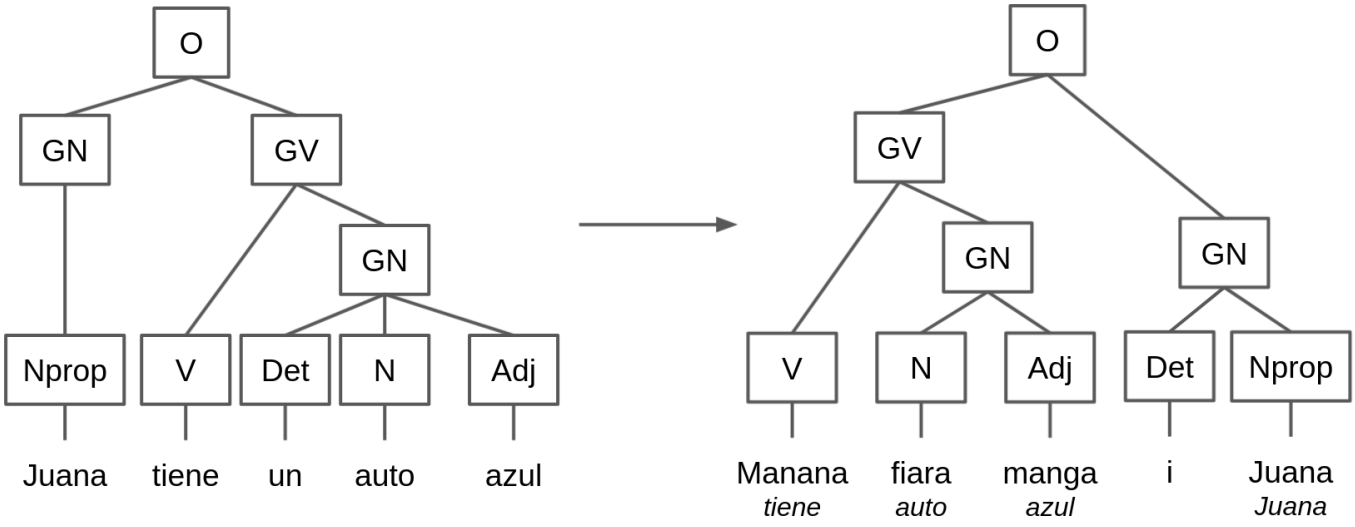
= (λQ . ∀x . verde(x) ∧ pájaro(x) → Q(x)) (λy . habla(y)) *[CAMBIO DE VARIABLE]*

= ∀x . verde(x) ∧ pájaro(x) → (λy . habla(y))(x) *[APLICACIÓN FUNCIONAL]*

= ∀x . verde(x) ∧ pájaro(x) → habla(x) *[APLICACIÓN FUNCIONAL]*

Ejercicio 4 [8 puntos]

a) Suponga que se quiere construir un sistema de traducción automática del español al malgache basado en transferencia sintáctica. Escriba las reglas de transferencia para dicho sistema, basándose en la siguiente traducción de ejemplo:



O → GN GV / O → GV GN
GN → Nprop / GN → Det Nprop
GN → Det N Adj / GN → N Adj

b) Utilizando el algoritmo de programación dinámica, calcule la distancia de Levenshtein entre las palabras ARBOL y MARMOL. Una vez calculada, identifique el camino para obtenerla.

	#	A	R	B	O	L
#	0	1	2	3	4	5
M	1	2	3	4	5	6
A	2	1	2	3	4	5
R	3	2	1	2	3	4
M	4	3	2	3	4	5
O	5	4	3	4	3	4
L	6	5	4	5	4	3

La distancia de Levenshtein entre las 2 palabras ARBOL y MARMOL es **3**.

Ejercicio 5

a) ¿En qué consiste un esquema de clasificación supervisada? Explique la diferencia entre modelos de clasificación probabilistas generativos y discriminativos, y nombre un método como ejemplo para cada enfoque.

Un esquema de clasificación supervisada es un escenario de aprendizaje automático en el que se cuenta para el entrenamiento con un conjunto de entradas y sus respectivas salidas esperadas.

Los método generativos aprenden la probabilidad conjunta de las clases y las entradas, con la cual se obtienen las probabilidades de cada clase para una entrada arbitraria (ej. Naïve Bayes). Por otro lado, los métodos discriminativos aprenden directamente la probabilidades condicionadas de las clases respecto a la entrada (ej. Regresión Logística).

b) Defina brevemente qué entiende por análisis de sentimiento. Mencione al menos dos tareas de PLN que se utilizan en esa tarea.

Es el proceso de detectar y extraer mediante el uso de técnicas de aprendizaje automático y procesamiento de lenguaje natural la connotación positiva o negativa de un texto en general o de la valoración subjetiva del emisor respecto a un tema.

En cuanto a las tareas de PLN involucradas, se pueden mencionar por ejemplo: *tokenizar* (separar en tokens el texto) y *normalizar* (llevar las palabras a algún formato estándar)

c) ¿Sobre qué objetivo o tarea está definido skip-gram con negative sampling? ¿Cómo se forman los ejemplos negativos en skip-gram con negative sampling?

Skip-gram con negative sampling está definido sobre la tarea de distinguir, para una palabra objetivo **w**, si una palabra **x** pertenece al contexto de **w** o no.

Cada ejemplo negativo se forma asociando una palabra aleatoria a la palabra objetivo.