Introducción al Procesamiento de Lenguaje Natural

Noviembre 2022 - Solución

Consideraciones generales	
i)	La prueba es sin material escrito.
ii)	Escriba nombre y C.I. en todas las hojas.
iii)	Numere todas las hojas.
iv)	En la primera hoja, indique el total de hojas.
v)	Comience cada ejercicio en una hoja nueva.
vi)	Utilice las hojas de un solo lado.
vii)	Entregue los ejercicios en orden.
viii)	El total de puntos es 70.

Ejercicio 1 [16 puntos]

Para cada afirmación diga si es Verdadera o Falsa. Justifique.

i) En la frase *La noche de las luces* la palabra ***de*** tiene categoría adverbio.

ii) Algo bueno que tiene representar documentos mediante el centroide de los word embeddings de sus palabras es que se captura la importancia del orden de las palabras en la oración.

iii) Skip-gram es uno de los algoritmos usados para construir word embeddings.

iv) Un modelo de lenguaje de n-gramas permite calcular la probabilidad de una secuencia de palabras.

v) El *accuracy* (o exactitud) es una buena medida para la performance de un clasificador binario en un corpus muy desbalanceado.

vi) La medida tf-idf sirve para medir qué tan importante es un término en un documento en el Modelo Booleano de Recuperación de Información.

vii) Para la construcción de corpus paralelos se pueden definir distintos tipos de alineación, como por ejemplo por documento o por oración.

viii) La medida BLEU tiene como principal objetivo dar la altura máxima de un árbol de derivación para una gramática libre de contexto.

Solución:

i) **Falso**. ***de*** es una preposición, en particular una preposición de lugar.

ii) **Falso.** Una de las desventajas que tiene el centroide es que calcula el vector promedio, dándole la misma importancia a todos los vectores involucrados.

iii) **Verdadero.** Es el visto en la clase teórica; existen otros.

iv) **Verdadero.** La probabilidad de la secuencia total se obtiene como producto de la probabilidad de los n-gramas.

v) **Falso.** En un corpus desbalanceado es mejor medir la performance utilizando la medida-F.

vi) **Falso.** La medida tf-idf sirve, por ejemplo, para asignar pesos a cada término de cada documento (cuan relevante es) de una colección de documentos y que se usa en el Modelo Vectorial. En el Modelo Booleano no se consideran las frecuencias de los términos en los documentos. También puede usarse por motores de búsqueda para marcar la relevancia de los documentos recuperados en la lista de resultados a partir de una consulta.

vii) **Verdadero.** Esa son dos de las posibles formas de alinear un corpus paralelo. También pueden ser alineados a nivel de palabra.

viii) **Falso.** La medida BLEU utiliza conteo de n-gramas para medir qué tanto se parecen dos secuencias de palabras.

Ejercicio 2 [20 puntos]

a) Dadas las siguientes oraciones:

Los perros comen huesos.
Los perros comen.
Un perro come lentamente.

i. Utilizando la notación bracket (corchetes rectos) segmente cada oración en sintagmas e indique: núcleo de cada sintagma, categoría de cada núcleo, categoría de cada sintagma.

ii. Escriba una Gramática Libre de Contexto G:(V,T,P,O) que las genere.

iii. Dé un ejemplo – por medio de una derivación – de una oración agramatical desde el punto de vista del español, pero que sea permitida por la gramática construida en la parte ii. Justifique.

Solución:

i.

*[Los **perros**]$_{GN}$ [**comen** [**huesos**]$_{GN}$]$_{GV}$.*
perros y huesos son Nombre (Sustantivo) comen (verbo)

*[Los **perros**]$_{GN}$ [**comen**]$_{GV}$.*
perros Nombre (Sustantivo) comen (verbo)

*[Un **perro**]$_{GN}$ [**come** [**lentamente**] $_{GAdv}$]$_{GV}$.*
perro Nombre (Sustantivo) come (verbo) lentamente (adverbio)

ii.

O → GN GV
GN → Det Nom | Nom
GV → V | V SAdv
Det → los | un
Nom → perros | perro | huesos
V → come | comen
GAdv → Adv
Adv → lentamente

iii. O ⇒ GN GV ⇒ Det Nom GV ⇒ un Nom GV ⇒ un huesos GV ⇒ un huesos V ⇒ un huesos come

Es agramatical porque no hay concordancia entre el número del determinante (un) con el nombre (huesos), ya que este último está en plural.

b) Considere la siguiente gramática libre de contexto:

O → GN GV | GV
GN → Nom | Det Nom | GN GP
GV → V GN | V GN GP | V
GP → Prep GN
Det → una | un | el | la

Prep → con | de
Nom → María | amigos | comienzo | final
V → mirando | mira | miro

Aplique el algoritmo CKY justificando su razonamiento a partir de la gramática G para cada una de las siguientes entradas indicando qué salida devuelve en cada caso:

1) *"Mirando la final con amigos"*
2) *"Mirando con amigos"*

Solución:

Como primer paso, se lleva la gramática a la Forma Normal de Chomsky

O → GN GV | mirando | mira | miro | V GN | Aux1 GP

GN → María | amigos | comienzo | final | Det Nom | GN GP

GV → mirando | mira | miro | V GN | Aux1 GP

Aux1→ V GN

GP → Prep GN

Det → una | un | el | la

Prep → con | de

Nom → María | amigos | comienzo | final

V → mirando | mira | miro

Con esta gramática G', se analizan las dos entradas:

i. *Mirando la final con amigos*

	Mirando (1)	la (2)	final (3)	con (4)	amigos (5)
0	**O GV V**		**GV Aux1 O**		**GV O Aux1**
1		**Det**	**GN**		**GN**
2			**Nom GN**		**GN**
3				**Prep**	**GP**
4					**Nom GN**

ii. Mirando con amigos

	Mirando (1)	con (2)	amigos (3)
0	**O GV V**		
1		**Prep**	**GP**
2			**Nom GN**

El algoritmo CKY es un reconocedor de oraciones válidas de acuerdo a una gramática.

En estos casos, para la entrada i. el algoritmo devuelve ***TRUE*** porque en la celda V_{05} está la variable O (símbolo inicial de la gramática).

Para el caso de la entrada ii. devuelve ***FALSE*** ya que no aparece la variable O en V_{03}.

Ejercicio 3 [20 puntos]

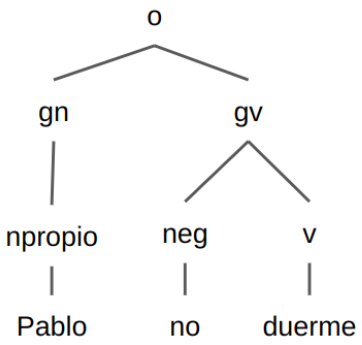
a) Sea la siguiente gramática con anotaciones semánticas:

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
gv → neg v	gv.sem = neg.sem(v.sem)
nom → n	nom.sem = n.sem
det → un	det.sem = λP. λQ. ∃x P(x) ∧ Q(x)
neg → no	neg.sem = λP . λx . ¬P(x)
n → pájaro	n.sem = λx.pájaro(x)
npropio → Pablo	npropio.sem = λP.P(pablo)
v → habla	v.sem = λx. habla(x)
v → duerme	v.sem = λx. duerme(x)
adj → simple	adj.sem = λP.λx. simple(x) ∧ P(x)

Utilizando las reglas anteriores dibuje el árbol sintáctico y derive la expresión lógica asociada a la oración:

Pablo no duerme

Solución:



o.sem

= gn.sem(gv.sem) [sustitución]

= λP.P(pablo)(neg.sem(v.sem)) [sustitución]

= λP.P(pablo)(λP.λx.¬P(x)(λx.duerme(x))) [sustitución]

= λP.P(pablo)(λP.λx.¬P(x)(λy.duerme(y))) [cambio de variable]

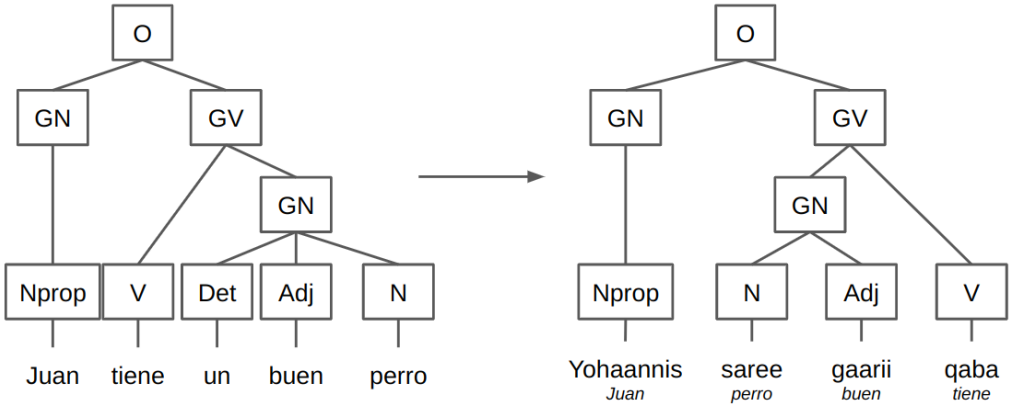
= λP.P(pablo)(λx.¬(λy.duerme(y))(x)) [aplicación funcional]

= λP.P(pablo)(λx.¬duerme(x)) [aplicación funcional]

= (λx.¬duerme(x))(pablo) [aplicación funcional]

= ¬duerme(pablo) [aplicación funcional]

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



Solución:

GV → V GN / GV → GN V

GN → Det Adj N / GN → N Adj

Ejercicio 4 [14 puntos]

a) Se tiene un conjunto de noticias clasificadas en una de tres categorías (deportes, policial, economía) y los resultados de predicción de un sistema automático para esas tres categorías.

Noticia	**Clase original**	**Predicción**
n1	Deportes	Policial
n2	Deportes	Deportes
n3	Deportes	Deportes
n4	Policial	Policial
n5	Policial	Economía
n6	Deportes	Deportes
n7	Economía	Deportes
n8	Deportes	Economía
n9	Deportes	Policial
n10	Economía	Deportes

Construya la matriz de confusión. Calcule el accuracy total, y los valores de precisión, recall y medida-F por cada clase.

b) Sea el siguiente texto de una noticia:

> *Una turba enojada casi lincha a una turista que subió sin permiso los escalones del Castillo de Kukulcán, una de las nuevas siete maravillas del mundo moderno que se encuentra en la zona arqueológica de Chichén Itzá, al sureste de México.*
>
> *El Instituto Nacional de Antropología e Historia (INAH) prohibió subirse al edificio sagrado de los mayas desde 2008, instaló un cordón de seguridad alrededor y anunció multas que van desde 50.000 (unos 2.558 dólares) a 100.000 pesos mexicanos (cerca de 5.115 dólares), dependiendo del daño que se cause a la estructura.*
>
> *El director general del INAH, Diego Prieto, aún no brindan informes del incidente, por lo que la multitud sigue exigiendo cárcel y la expulsión de Yucatán, "y, si es del extranjero, que se vaya de México", gritaban los presentes.*
>
> *En incidentes anteriores, el INAH informó que sí se castigaría en acuerdo con el Ministerio de Turismo, quien establece las penas y sanciones contra aquellas personas que dañen o exploten monumentos arqueológicos inmuebles sin autorización del Instituto.*

En un proceso de extracción de información sobre el texto anterior dé:

i. al menos 3 entidades con nombre que pertenezcan distintas categorías.

ii. al menos 2 relaciones binarias indicando sus argumentos.

Solución:

a) La clase esperada se ve por filas, la clase predicha se ve por columnas.

	D	P	E
D	3	2	1
P	0	1	1
E	2	0	0

acc = 4/10 = 0.4

$p_d = 3/5 = 0.6$	$r_d = 3/6 = 0.5$	$f_d = 6/11 = 0.545$
$p_p = 1/3 = 0.333$	$r_p = 1/2 = 0.5$	$f_p = 2/5 = 0.4$
$p_e = 0$	$r_e = 0$	$f_e = 0$

b)

Entidades con nombre (al menos una de cada tipo):

Lugares:*Castillo de Kukulcán, Chichén Itzá, México, Yucatán…..*

Organizaciones: *El Instituto Nacional de Antropología e Historia (INAH), Ministerio de Turismo*

Personas: *Diego Prieto*

Relaciones:

- linchar <una turba enojada, una turista>
- subir_a_la_cima <la turista, majestuoso edificio>
- subir_sin_permiso <una turista, los escalones del Castillo de Kukulcan>

(estas son algunas….)