Introducción al Procesamiento de Lenguaje Natural

Diciembre 2020 - Soluciones

Ejercicio 1 [puntos]

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

i. HMM es un método discriminativo de clasificación secuencial.

Falso. HMM es generativo.

ii. Un modelo de trigramas intenta predecir cada palabra a partir de las 2 anteriores.

Verdadero.

iii. La frase "muy buenos estudiantes" es más probable que la frase "buenos muy estudiantes" según un modelo de unigramas.

Falso. Tienen la misma probabilidad.

iv. La técnica de suavizado de Laplace para bigramas consiste en sumar uno en el numerador y $|V|^2$ en el denominador del cálculo de las probabilidades de una palabra dado el contexto.

Falso. Se suma 1 en el numerador y $|V|$ en el denominador.

v. La distancia de mínima edición es una técnica que utiliza las probabilidades para diferenciar un lema de otro.

Falso. No es una técnica probabilística.

b) Para los siguientes ejemplos, diga si la afirmación es Verdadera o Falsa

i. *El joven es alto*

La palabra "joven" tiene categoría adjetivo. FALSO

ii. *El hombre es joven*

La palabra "joven" tiene categoría adjetivo. VERDADERO

iii. *Este joven es alto*

La palabra "este" tiene categoría adverbio. FALSO

iv. *El joven es famoso*

La palabra "el" tiene categoría determinante. VERDADERO

v. *El joven es muy famoso*

La palabra "muy" tiene categoría preposición. FALSO

Ejercicio 2

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

O → GN GV

GN → Det Nom | Nom | GN GP

GV → V | V GN | V GN GP | V GP

GP → Prep GN

Det → una | un | el | la

Nom → Juan | María | TV | fuente

V → mira | sale | come

Prep → para | de

a) Aplique el algoritmo CKY para la entrada "*Juan come de la fuente*" a partir de la gramática. ¿Qué salida devuelve el algoritmo?

Primero para aplicar CKY la gramática debe estar en Forma Normal de Chomsky (Libre de Contexto, simplificada y con reglas del tipo: A → BC o A → a ; siendo A,B,C Variables y a un terminal)

O → GN GV

GN → Det Nom | GN GP | Juan | María | TV | fuente

GV → V GN | V GP | V Aux1 | mira | sale | come

GP → Prep GN

Aux1 → GN GP

Det → una | un | el | la

Nom → Juan | María | TV | fuente

V → mira | sale | come

Prep → para | de

	Juan (1)	come (2)	de (3)	la (4)	fuente (5)

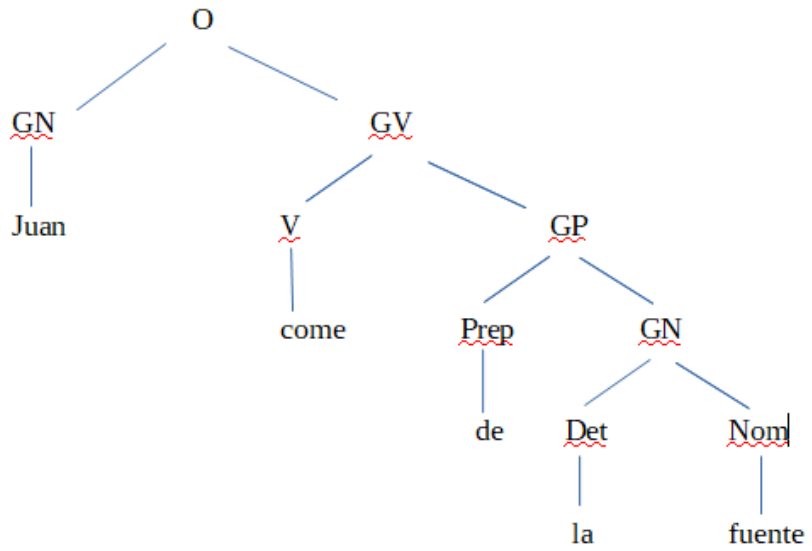
0	GN Nom	O	Ø	Ø	O
1		GV V	Ø	Ø	GV
2			Prep	Ø	GP
3				Det	GN
4					Nom GN

De donde como en la casilla (0,5) contiene O, el algoritmo devuelve TRUE.

b) El algoritmo CKY es un reconocedor de oraciones válidas. ¿Qué modificaciones se debieran hacer para que además devuelva el o los árboles de análisis sintáctico encontrados? Dibuje el o los árboles sintácticos posibles para la oración de la parte a).

Debiera guardar en cada paso (casilla) para cada variable, las variables (y casillas) de donde proviene; es decir, las variables del lado derecha de la regla que se aplica en cada iteración del algoritmo.

En este caso, hay un sólo árbol, se dibuja a continuación:



c) ¿Cómo podría modificarse el algoritmo para que asocie una probabilidad a cada árbol sintáctico válido?

La idea sería intentar construir una gramática libre de contexto probabilista. Para eso es necesario que surja a partir de un corpus, donde se le agregan probabilidades a las reglas de manera tal que exista para cada variable del lado izquierdo, una distribución de probabilidades.

De esa manera, se puede calcular la probabilidad de una oración (multiplicando las probabilidades asociadas a cada regla empleada en la construcción de árbol), y en caso de que exista más de un árbol, sumando cada una de sus probabilidades.

Ejercicio 3 [puntos]

gn → det nom	gn.sem = det.sem(nom.sem)
nom → n	nom.sem = n.sem
nom → nom adj	nom.sem = adj.sem(nom.sem)
nom → nom pp	nom.sem = pp.sem(nom.sem)
pp → prep npropio	prep.sem(npropio.sem)
prep → de	prep.sem = λx. λP. λy.de(x,y) ∧ P(y)
det → un	det.sem = λP. λQ. ∃x P(x) ∧ Q(x)
n → libro	n.sem = λx.libro(x)
npropio → Sartre	npropio.sem = Sartre
adj → interesante	adj.sem = λP.λx. interesante(x) ∧ P(x)

Utilizando las reglas anteriores realice una derivación para calcular la representación semántica del grupo nominal: ***un libro interesante de Sartre***

NÚM	REGLA	SEMÁNTICA
1	gn → det nom	gn.sem = det.sem(nom.sem)
2	nom → nom adj	nom.sem = adj.sem(nom.sem)
3	nom → n	nom.sem = n.sem = λx.li→bro(x)
4	n → libro	n.sem = λx.libro(x)
5	adj → interesante	adj.sem = λP.λx. interesante(x) ∧ P(x)
6	2 → 3 , 5	nom.sem = (λP.λx. interesante(x) ∧ P(x)) λz.libro(z) ⊳ λx. interesante(x) ∧ λz.libro(z)(x) ⊳ λx. interesante(x) ∧ libro(x)
7	nom → nom pp	nom.sem = pp.sem (nom.sem)
8	pp → prep npropio	pp.sem = prep.sem (npropio.sem)
9	prep → de	prep.sem = λx. λP. λy.de(x,y) ∧ P(y)
10	npropio –> Sartre	npropio.sem = sartre
11	8 → 9 , 10	nom.sem = λx. λP. λy.de(x,y) ∧ P(y) (sartre) ⊳ λP. λy.de(sartre,y) ∧ P(y)
12	7 → 6 , 8	nom.sem = λP. λy.de(sartre,y) ∧ P(y) (λz. interesante(z) ∧ libro(z)) ⊳ λy.de(sartre,y) ∧ (λz. interesante(z) ∧ libro(z))(y) ⊳ λy. de(sartre,y) ∧ interesante(y) ∧ libro(y)
13	det → un	det.sem = λP. λQ. ∃x P(x) ∧ Q(x)
14	1 → 13 , 12	gn.sem = λP. λQ. ∃x P(x) ∧ Q(x) (λy. de(sartre,y) ∧ interesante(y) ∧ libro(y)) ⊳ λQ. ∃x (λy. de(sartre,y) ∧ interesante(y) ∧ libro(y))(x) ∧ Q(x) λQ. ∃x de(sartre,x) ∧ interesante(x) ∧ libro(x) ∧ Q(x)

Ejercicio 4

a) Considere el siguiente texto de una noticia:

La intendenta de Montevideo, Carolina Cosse, tuvo este jueves en la plaza Las Pioneras su ceremonia de asunción, en donde se dio el traspaso de mando con el ahora ex-intendente Christian Di Candia.
Luego de una extensa muestra artística de la Orquesta Filarmónica de Montevideo, que contó con la participación de los músicos Ruben Rada, Eli Almic, Christian Cari y Cristina Fernández, la intendenta firmó su acta de asunción y luego dio un discurso en el que reconoció que la actualidad "está difícil", pero afirmó que "siempre es hora de comenzar".

Encuentre:

i) 3 entidades con nombre que pertenezcan a categorías distintas

Por ejemplo:

Carolina Cosse: **Persona**

Orquesta Filarmónica de Montevideo: **Organización**

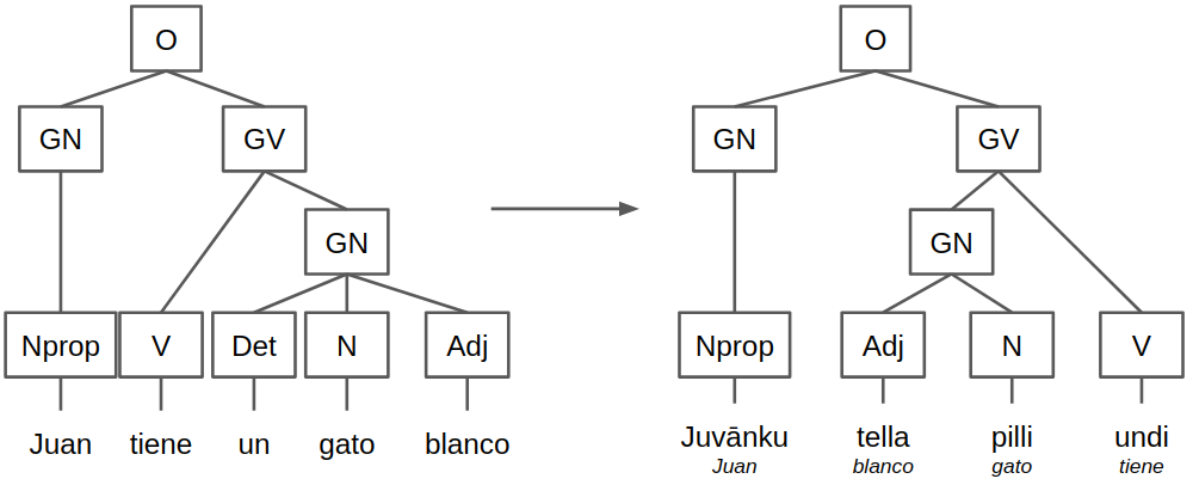
Plaza Las Pioneras: **Lugar**

ii) 2 coreferencias

Por ejemplo:

Carolina Cosse (párrafo 1) que tiene las siguientes coreferencias (párrafo 2): firmó su acta de asunción; dio un discurso; afirmó que es hora de comenzar

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



Reglas:

GV → V GN / GV → GN V

GN → Det N Adj / GN → Adj N

Diccionario:

Nprop → Juan / Nprop → Juvānku

V → tiene / V → undi

N → gato / N → pilli

Adj → blanco / Adj → tella

c) i) ¿Qué tipo de problema es, desde el punto de vista computacional, el de asignar un idioma a un documento?
ii) Teniendo un corpus de 1000 documentos, cada uno anotado con su idioma (español o inglés), donde 975 son en inglés. ¿Existe algún problema con este corpus? En caso afirmativo, ¿cómo puede resolverlo? Justifique brevemente sus respuestas.

i) Es un problema de clasificación (se le asigna una categoría de un conjunto a cada instancia).

ii) El problema es que la precisión es muy alta porque hay muy pocos ejemplos en un solo idioma, por lo que la precisión va a dar muy alta (975/1000). Este corpus está desbalanceado, deberíamos agregar más instancias de documentos en español.

Ejercicio 5

Para cada frase, elegir la opción correcta:

1. **El algoritmo de Lesk de desambiguación semántica**

d. compara las palabras del contexto de la palabra a desambiguar con las glosas de las distintas acepciones.

2. **El método Naïve Bayes realiza las siguientes simplificaciones**

a. considera que la probabilidad conjunta de un conjunto finito de variables aleatorias que representan valores de atributos es el producto de las probabilidades

3. **Los word embeddings son**

b. una representación vectorial para las palabras a partir de los contextos donde ocurren en un corpus.

4. **Skip-gram con negative sampling está formulado en base a la siguiente tarea:**

b. Distinguir para cada palabra, a las palabras que aparecen en su contexto de otras que no.

5. **Los métodos discriminativos de clasificación probabilista**

d. para cada entrada generan una distribución de probabilidad sobre las clases.