Introducción al Procesamiento de Lenguaje Natural

Diciembre 2020

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 [15 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.

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

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

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.

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

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

i. *El joven es alto*

La palabra "joven" tiene categoría adjetivo.

ii. *El hombre es joven*

La palabra "joven" tiene categoría adjetivo.

iii. *Este joven es alto*

La palabra "este" tiene categoría adverbio.

iv. *El joven es famoso*

La palabra "el" tiene categoría determinante.

v. *El joven es muy famoso*

La palabra "muy" tiene categoría preposición.

Ejercicio 2 [20 puntos]

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

Ejercicio 3 [15 puntos]

Sea la siguiente gramática con anotaciones semánticas:

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	pp.sem = prep.sem(npropio.sem)
prep → de	prep.sem = $\lambda x.\ \lambda P.\ \lambda y.de(x,y) \wedge P(y)$
det → un	det.sem = $\lambda P.\ \lambda Q.\ \exists x\ P(x) \wedge Q(x)$
n → libro	n.sem = $\lambda x.$libro(x)
npropio → Sartre	npropio.sem = sartre
adj → interesante	adj.sem = $\lambda P.\lambda x.$ interesante(x) $\wedge$ P(x)

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

Ejercicio 4 [10 puntos]

a) Considere el siguiente texto de una noticia:

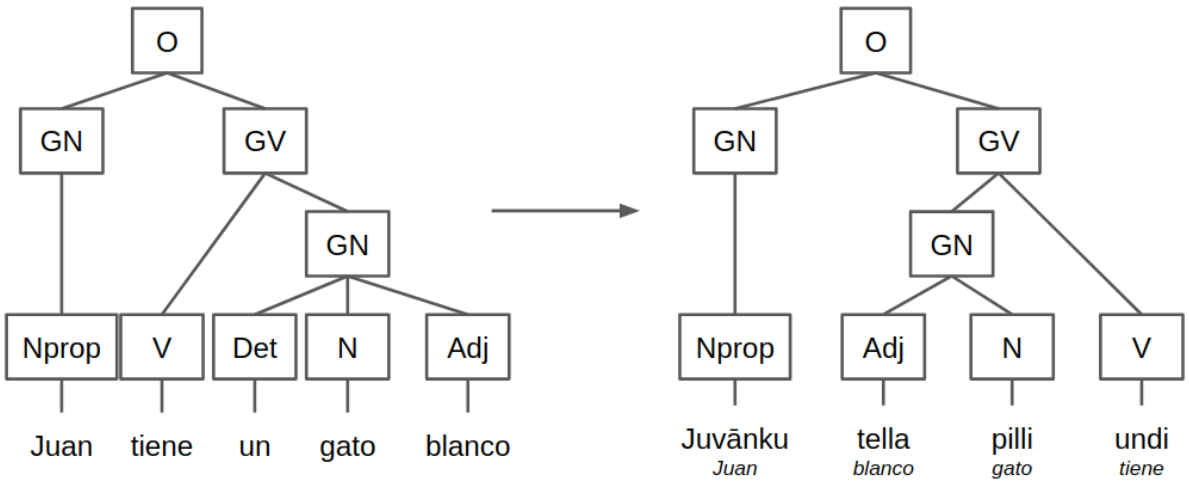
La intendenta de Montevideo, Carolina Cosse, tuvo este jueves en la plaza Las Pioneras su ceremonia de asunción, en donde se dio el traspaso de mando con el ahora ex-intendente Christian Di Candia.
Luego de una extensa muestra artística de la Orquesta Filarmónica de Montevideo, que contó con la participación 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

ii) 2 coreferencias

b) 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, a partir de la siguiente traducción de ejemplo:



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 9750 son en inglés. ¿Existe algún problema con este corpus? En caso afirmativo, ¿cómo puede resolverlo? Justifique brevemente sus respuestas.

Ejercicio 5 [10 puntos]

Para cada frase, elegir la opción correcta:

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

a) se basa en búsquedas en la Wikipedia.

b) compara los synsets de c/u de las distintas acepciones en conflicto con los hiperónimos de las otras.

c) calcula distancia de los embeddings de las palabras del contexto de la palabra a desambiguar con los embeddings de los synsets de cada acepción.

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.

b) considera equiprobables a todos los posibles valores de los atributos.

c) excluye del modelo los valores de los atributos con probabilidad < 0.1

d) realiza una selección aleatoria de atributos a considerar, de modo que la cantidad total de atributos no supere un parámetro P.

3. **Los word embeddings son**

a) una representación one-hot para las palabras.

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

c) una forma de Bag of Words con tf-idf.

d) una representación vectorial para las palabras basada en su morfología.

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

a) Predecir una palabra a partir de su contexto.

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

c) Predecir la próxima palabra a partir del contexto previo.

d) Obtener la probabilidad de la oración.

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

a) discriminan la clase de la entrada más probable.

b) generan una distribución de probabilidad para las entradas.

c) son aplicables cuando se tienen las probabilidades de transición entre las clases.

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