Introducción al Procesamiento de Lenguaje Natural

Diciembre de 2007

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 80

Ejercicio 1 [16 puntos]

Para el texto que se da a continuación, responda las preguntas que se plantean.

.....
El vendedor era un muchacho correntino, bajo y de pelo cortado al rape, que usaba siempre botines amarillos. **El otro**, encargado de los libros, era un hombre hecho ya, muy flaco y de cara color paja. Creo que nunca **lo** vi reírse, mudo y contraído en su Mayor con estricta prolijidad de rayas y tinta colorada. Se llamaba Figueroa; era de Catamarca.
Ambos, comenzando por salir juntos, trabaron estrecha amistad y, como ninguno tenía familia en Laboulaye, habían alquilado un caserón con sombríos corredores de bóveda, obra de un escribano que murió loco allá.
Los dos primeros años no tuvimos la menor queja de nuestros hombres. Poco después comenzaron, cada uno a su modo, a cambiar de modo de ser.
El vendedor —se llamaba Tomás Aquino— llegó cierta mañana a la barraca con una verbosidad exuberante. Hablaba y reía sin cesar, buscando constantemente no sé qué en los bolsillos. Así estuvo dos días. Al tercero cayó **[**con un fuerte ataque de gripe**]**; pero volvió después de almorzar, inesperadamente curado. Esa misma tarde, Figueroa tuvo que retirarse con desesperantes estornudos preliminares que lo habían invadido de golpe. Pero todo pasó en horas, a pesar de los síntomas dramáticos. Poco después se repitió lo mismo, y así, por un mes: la charla delirante de Aquino, los estornudos de Figueroa, y cada dos días un fulminante y frustrado ataque de gripe.
Esto era lo curioso. Les aconsejé que se hicieran examinar atentamente, pues no se podía seguir así. Por suerte todo pasó, regresando ambos a la antigua y tranquila normalidad, el vendedor entre las tablas, y Figueroa con su pluma gótica.
.....
Fragmento de *Las rayas* de Horacio Quiroga

a) Decir cuál de los dos sintagmas nominales en negrita (*El vendedor* y *El otro*) es el antecedente del elemento anafórico *lo* (en negrita y subrayado). ¿En qué criterio se basó para identificarlo?
Solución:
El antecedente es **el otro.** Podemos aplicar el criterio según el cual las entidades mencionadas más recientemente en el discurso tienen mayor probabilidad de ser antecedentes de elementos anafóricos, siempre que haya concordancia (en este caso, los dos sintagmas propuestos concuerdan en género y número con el pronombre **lo**, por lo cual el criterio de concordancia no permite decidir entre los dos).

b) Para la oración *Ambos, ... allá* (marcada en negrita), decir cuáles son los verbos conjugados e identificar las proposiciones que éstos determinan. Mostrar cómo se coordinan o subordinan las diferentes proposiciones, utilizando corchetes.
Solución:
[1 [2 Ambos, comenzando por salir juntos, trabaron estrecha amistad 2] y, [3 [4 como ninguno tenía familia en Laboulaye 4] , habían alquilado un caserón con sombríos corredores de bóveda, obra de un escribano [5 que murió loco allá 5] 3] 1].

1 es una coordinación entre 2 y 3, el elemento coordinante es *y*
2 tiene como núcleo *trabaron*
3 tiene como núcleo *habían*
4 está subordinada a 3 por el elemento *como* y tiene como núcleo *tenía*
5 está subordinada a 3 por el elemento *que* y tiene como núcleo *murió*

c) Indicar cuál es el sujeto y cuál es el complemento indirecto de la oración (simplificada) *Les aconsejé ... atentamente*, marcada en el texto con subrayado. Justifique.
Solución:
Les aconsejé que se hicieran examinar atentamente.

sujeto: primera persona del singular (sujeto omitido *yo*, se recupera a partir de la termiación del verbo).
justificación: si cambio la conjugación verbal (les aconsejamos que se hicieran...., les aconsejó que se hicieran ...) ningún otro elemento de la oración cambia.

objeto indirecto: *les*.
justificación: *les* es uno de los clíticos indirectos, es decir, que se ponen en lugar del objeto indirecto.

d) Analizar el grupo de palabras *con un fuerte ataque de gripe*, marcado en el texto con corchetes:

i. ¿Qué tipo de grupo o sintagma es?
Solución:
Es un sintagma preposicional.

ii. ¿Cuál es el elemento principal del sintagama, es decir, el que le da la categoría?
Solución:
La preposición *con*.

iii. ¿Qué sintagmas más pequeños contiene? Dar el tipo y la palabra principal de cada sintagma identificado.
[sp **con** [sn un fuerte **ataque** [sp **de** [sn **gripe**]]]]

Ejercicio 2 [16 puntos]

Considere los **verbos** en el idioma ficticio Itio:

1 Los verbos en infinitivo tienen dos terminaciones: *tu* (como *itu*, que en español quiere decir *correr*) y *nu* (como *valnu* y *dinu*, que quieren respectivamente decir *saltar* y *meditar en medio del caos*).

2. Hay tres personas en Itio, y no se distingue plural de singular.

3. Para la conjugación en presente, a los verbos se les agregan los siguientes sufijos:

Terminación	Persona	Sufijo	Ejemplo
tu	1era	-fu	itfu
tu	2da	-fa	itfa
tu	3era	-fe	itfe
nu	1era	-nu	valnu
nu	2da	-no	valno
nu	3era	-ne	dine

Observar que al agregar el sufijo, los verbos terminados en *tu* pierden la *u*, y los terminados en *nu,* pierden toda la terminación.

4. Al conjugar un verbo en presente, puede (opcionalmente) agregarse un sufijo que denota modalidad: *natnat*, para denotar *con violencia*, *uz*, para denotar *rápidamente, y top*, para denotar *sonriendo misteriosamente.* Por ejemplo, *valnunatnat* quiere decir *salto violentamente.*

5. Por cuestiones ortográficas, en Itio no están permitidas las vocales dobles. Cuando aparezcan al combinar morfemas, se convertirá en una sola.

a) Describa brevemente los tres componentes principales de un analizador morfológico. Describa cada componente para un hipotético analizador morfológico para los verbos en Itio.

Los tres componentes principales son: ***Lexicón****: conjunto de morfemas del lenguaje (raíces y afijos).* ***Reglas morfotácticas****: reglas para el ordenamiento de los morfemas.* ***Reglas ortográficas****: modificaciones en la forma de las palabras al combinarse los morfemas. En Itio el lexicón está dado por el conjunto formado por las raices y los sufijos, las morfotáticas son las reglas que dicen que los sufijos van después de las raíces, y la única regla ortográfica es la del punto 5.*

b) Construya una expresión regular utilizando el álgebra de Xerox, que relacione un verbo en Itio conjugado en presente con su análisis morfológico. Utilice las siguientes marcas léxicas para la salida: +*Verb* (Raíz), +*Primera*(Primera persona), +*Segunda*(Segunda Persona),+*Tercera*(Tercera Persona), +Violencia, +Rapido, +SonriendoMisteriosamente. Suponga que las únicas raíces verbales son las que se presentaron en los ejemplos (*itu, valnu, dinu)*, pero diseñe la solución de modo que pueda ampliarse a otras raíces fácilmente.

Ejemplos:

dinuz	dinu+Verb+Primera+Rapido
itfatop	itu+Verb+Segunda+SonriendoMisteriosamente
valno	valnu+Verb+Segunda

Fuente xfst:

```
define roottu [{itu}:{it}] %+Verb:0;
define rootnu [{dinu}:{di}  | {valnu}:{val}] %+Verb:0;
define conjugaciontu [roottu [%+Primera:{fu} | %+Segunda:{fa} | %+Tercera:{fe}]];
define conjugacionnu [rootnu [%+Primera:{nu} | %+Segunda:{no} | %+Tercera:{ne}]];
define conjugacion conjugacionnu | conjugaciontu;
define sufijo %+Violencia:{natnat} | %+Rapido:{uz} | %+SonriendoMisteriosamente:{top};
define ortografia [u u] -> u ;
define itiov [conjugacion (sufijo)] .o. ortografia;
```

Ejercicio 3 [6 puntos]

Suponga que utiliza un Modelo de Markov de Estados Ocultos (HMM) para hacer POS tagging.

a) Los HMMs son un caso particular de *inferencia bayesiana*. Aplique la fórmula de Bayes para mostrar cómo estima este método la probabilidad de una secuencia de tags, dada una secuencia de palabras.

Lo que se busca es la secuencia de tags T que maximice la probabilidad P(T |O), dada una secuencia O de observaciones. Por la regla de Bayes

$$P(T|O) = \frac{P(T|O)P(O)}{P(T)}$$

Por lo tanto buscamos argmax P(T|O)P(O). Ambas probabilidades pueden estimarse calculando frecuencias de aparición en un corpus etiquetado.

b) ¿Qué hipótesis principales asumen los HMMs para simplificar el cálculo de la fórmula obtenida en el punto anterior?

Como el cálculo directo de la probabilidad anterior es muy costoso, los HMMs asumen que:

- *la probabilidad de que aparezca una palabra sólo depende de su etiqueta*
- *la probabilidad de que apareza una etiqueta sólo depende de la etiqueta anterior(hipótesis de bigrama)*

Ejercicio 4 [18 puntos]

a) Describa brevemente las ventajas y desventajas de aplicar estrategias puramente *top-down* o *bottom-up.*

La principal ventaja de los analizadores *bottom-up* es que su análisis se encuentra orientado por la entrada: no exploran análisis que no sean consistentes con la entrada. Sin embargo, esto tiene como problema que se pueden generar sub-análisis que no sirven para alcanzar al símbolo inicial de la gramática.

En cambio, los analizadores top-down generan árboles que comienzan con el símbolo inicial, pero "pierden tiempo" explorando análisis que no son consistentes con la entra que están analizando.

b) Considere la siguiente gramática en Forma Normal de Chomsky.

O → GV GN | GN GV
GV → V GN
GN → Det N | Det GNP
GNP → N GP
GP → Prep GN
N → sala | masa | hombre
Det → el | la | los
V → sala | trae
Prep → con | de

Aplique el algoritmo CYK para la entrada "*El hombre sala la masa de la sala*".

O							
		GV					

O			GN				
				GNP			
		GV			GP		
GN			GN			GN	
Det	N	N, V	Det	N	Prep	Det	N, V
El	*hombre*	*sala*	*la*	*masa*	*de*	*la*	*sala*

Ejercicio 5 [16 puntos]

a. Se agregan 9 reglas, escritas a continuación de la gramática.

o → gn gv [o.sem = gv.sem (gn.sem)]
gn → det nominal [gn.sem = < det.sem x nominal.sem(x) >]
gn → npropio [gn.sem = npropio.sem]
nominal → n [nominal.sem = λx isa(x,n.sem)]
gv → v [gv.sem = v.sem]
gv → v gn [gv.sem = v.sem (gn.sem)]
v → lee [v.sem = λx λy ∃e leer (e) ∧ lee(e,y) ∧ leído(e,x)]
det → un [det.sem = ∃]
n → libro [n.sem = libro]
v → entrega
[v.sem = λx λy λz ∃e entregar (e) ∧ entrega(e,z) ∧ entregado(e,y) ∧ receptor(e,x)]

Se agregan las siguientes reglas :

1- gv → v gn1 a gn2 [gv.sem = (v.sem (gn2.sem)) (gn1.sem)]
2- gv → v gn1 prep-de gn2 [gv.sem = (v.sem (gn2.sem)) (gn1.sem)]

3- v → recibe
[v.sem = λx λy λz ∃e recibir (e) ∧ recibe(e,z) ∧ recibido(e,y) ∧ dador(e,x)]

4- v → recibe
[v.sem = λy λz ∃e recibir (e) ∧ recibe(e,z) ∧ recibido(e,y)]

5- prep-de → de [prep-de.sem = λx λy de(y,x)]
6- pp → prep gn [pp.sem = prep.sem(gn.sem)]
7- nominal → nominal pp [nominal.sem = λx nominal.sem (x) ∧ pp.sem(x)]

8- npropio → Juan [npropio.sem = juan]
9- npropio → Pedro [npropio.sem = pedro]

Notas:
En la regla 1, la preposición "a" simplemente se consume, no interviene en la representación semántica. De hecho, no aporta significado.
La ambigüedad en la oración "Juan recibe un libro de Pedro" corresponde a las siguientes dos parentizaciones posibles en el grupo verbal : recibe (un libro) (de Pedro) y recibe (un libro (de Pedro)) – En el primer caso se utiliza la regla 3 y en el segundo las reglas 4, 5, 6 y 7.
En las reglas 1 y 2 se usan las variantes gn1 y gn2 para mantener la referencia al consituyente adecuado en la representación semántica.

b. Utilizando las gramática completada realice la derivación de una representación semántica para la oración :

Juan entrega un libro a Pedro.

n → libro [libro]
libro
nominal → n [λx isa(x,libro)]
libro
gn → det nominal [< ∃ x isa (x,libro) >]
un libro
npropio → Pedro [pedro]
Pedro
gn → npropio [pedro]
Pedro
v → entrega
[λx λy λz ∃e entregar (e) ∧ entrega(e,z) ∧ entregado(e,y) ∧ receptor(e,x)]
entrega

gv → v gn1 a gn2
[(λx λy λz ∃e entregar (e) ∧ entrega(e,z) ∧ entregado(e,y) ∧ receptor(e,x) (pedro)) (< ∃ x isa (x,libro) >)] =
[λz∃e entregar(e)∧entrega(e,z)∧entregado(e, <∃ x isa(x,libro) >) ∧ receptor (e,pedro)]
entrega un libro a Pedro

o → gn gv [
[λz ∃e entregar(e)∧entrega(e,z)∧entregado(e, <∃ x isa(x,libro) >) ∧ receptor (e,pedro) (juan)] =

[∃e entregar(e)∧entrega(e,juan)∧entregado(e, <∃ x isa(x,libro) >) ∧receptor (e,pedro)]
=
[∃e∃x entregar(e)∧entrega(e,juan)∧entregado(e,x) ∧isa(x,libro) ∧receptor (e,pedro)]

Juan entrega un libro a Pedro

Ejercicio 6 [6 puntos]

a) Explique brevemente en qué se diferencian los llamados Sistemas de Recuperación de Información de los Sistemas de Información tradicionales (basados en bases de datos).

Por *Sistema de Información tradicional*, entendemos a los llamados Sistemas de Gestión de Bases de Datos, es decir, software dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que la utilizan. Se compone de un lenguaje de definición de datos, de un lenguaje de manipulación de datos y de un lenguaje de consulta. El propósito general de los sistemas de gestión de base de datos es el de manejar de manera clara, sencilla y ordenada un conjunto de datos.

Por otro lado están los Sistemas de Recuperación de Información, que son aquellos centrados en la búsqueda de información en documentos, búsqueda de los mismos documentos, la búsqueda de metadatos que describan documentos. Tienen como objetivo, dada una necesidad de información (consulta + perfil de usuario) y un conjunto de documentos, recuperar y devolver en forma ordenada de más a menos relevantes (con algún criterio), minimizando el tiempo empleado.
En BD, se sabe **exactamente** lo que se quiere, mientras que en IR, **no existe** la respuesta correcta.

En IR, cada documento puede ser más o menos relevante, y esto puede cambiar según el usuario y la situación.
En BD, importa la **eficiencia** (de tiempo de recuperación y de espacio de almacenamiento), mientras que en IR, importa quizás más la **calidad** de la respuesta

b) Mencione cuáles son las características principales del Modelo Vectorial, que lo diferencian sustancialmente de su antecesor, el Modelo Booleano.

La caracterización formal planteada por el Modelo Vectorial es que representa documentos y consultas a través de vectores en un espacio lineal multivectorial. Se trata de vectores de términos. Los rankings de relevancia de documentos en una búsqueda por palabras clave se pueden calcular, teniendo en cuenta la teoría de las semejanzas del documento, comparando la desviación de los ángulos de cada vector del documento y el vector original de la pregunta.
Un documento *di* se modela como un vector: $di = (wi1, wi2, \ldots, wik)$
donde, a cada término *j*, en un documento *di* se le asigna un peso *wij*

En el Modelo Booleano, no se tiene pesos, y los documentos se indizan por la ocurrencia o no de determinado término. No se consideran por lo tanto términos que parezcan frecuentemente en el documento, alcanza con que aparezca para ser seleccionado. Esto provoca que no exista el concepto de similitud – tal como lo proipone el modelo vectorial- entre documentos y consulta lo cual provoca que no pueda existir documentos ordenados por un cierto ranking (documentos más o menos relevantes). Da lo mismo que cumpla una o todas las cláusulas de un OR y no considera el "*casi* " de un documento (documento que cumpla casi todas las cláusulas de un AND).