Introducción al Procesamiento de Lenguaje Natural

SOLUCIÓN - Diciembre de 2006

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.

Ejercicio 1 [6 puntos]

a) Describa brevemente las dos hipótesis principales sobre la forma en que se realiza el procesamiento morfológico humano.

Existen dos hipótesis principales para el procesamiento morfológico (como se analiza la estructura de las palabras y como se generan las mismas) en humanos. La primera supone que se cuenta con una lista de todas las palabras posibles (sin considerar su estructura interna). La segunda, que las palabras se construyen a partir de una lista de morfemas (raíces y afijos) que se combinan de acuerdo a ciertas reglas de ordenamiento (morfotácticas).

b) Mencione dos ventajas de los transductores como herramienta para el desarrollo de analizadores morfológicos.

Por su propia definición, los transductores permite realizar el análisis léxico y la generación como procesos simétricos, por lo que resuelto el problema del análisis se resuelve el de la generación, y viceversa. Por sus propiedades, permiten realizar un procesamiento eficiente de la entrada (en aquellos casos en se pueden secuencializar, es O(n), siendo n el tamaño de la entrada). Los transductores pueden generarse fácilmente a partir de un álgebra de expresiones regulares que denota las relaciones regulares que ellos computan.

Ejercicio 2 [16 puntos]

Considere los **nombres** en el idioma ficticio Itio (ligeramente basado en el Esperanto), que se construyen según las siguientes reglas:

1. Comienzan con una raíz. Para este ejemplo consideraremos las siguientes: *hund, kat, bird, elefant (*en español: *perro*, *gato*, *pájaro*, *elefante* , respectivamente).
2. Una raíz por sí misma no es una palabra válida.
3. Todos los nombres tienen una sufijo *o.* Por lo tanto *elefanto* y *hundo* son palabras válidas.
4. Una palabra como *hundo* no tiene género marcado. Para marcarla como femenina, se agrega el sufijo *in* entre la raíz y el sufijo o. Por ejemplo *hundino* (perra) es una palabra válida.
5. Para marcar una palabra como diminutivo, se agrega un sufijo *et* entre la raíz y el sufijo *o*, por ejemplo *elefanteto* es una palabra válida. Para marcar un aumentativo, se agrega el sufijo *eg*.
6. Los sufijos femenino, diminutivo y aumentativo pueden co-ocurrir una cantidad arbitraria de veces.
7. Si una raíz termina en t, al agregarle el sufijo in, ésta se convierte en v, por cuestiones de pronunciación. Por lo tanto *gata* se escribe *kavino*.

a) Describa brevemente los tres componentes principales de un analizador morfológico. Describa cada componente para un hipotético analizador morfológico para los nombres 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 7.*

b) Construya una expresión regular utilizando el álgebra de Xerox, que relacione un nombre en Itio con su análisis morfológico. Utilice las siguientes marcas léxicas para la salida: +*Noun* (Raíz), +*Aug*(Aumentativo), +*Dim*(Diminutivo),+*Fem*(Femenino). Por ejemplo, a la forma léxica *elefant+Noun+Fem* le corresponderá la palabra *elefanvino*.

La expresión regular itio denota, en el álgebra de expresiones regulares de Xerox, la relación regular que mapea palabras y estructuras de los nombres en Itio (primero se definen las raíces y los afijos), luego se los combina de acuerdo a las reglas morfotáticas, para finalmente componer la relación con el reemplazo que denota los cambios ortográficos.:

```
define root [{hund}|{kat}|{bird}|{elefant}] %+Noun:0;

define sufijoo 0:o;

define diminutivo %+Dim:{et};

define aumentativo %+Aug:{eg};

define femenino %+Fem:{in};

define ortografia t -> v || _ {in};

define itio [root (femenino|diminutivo|aumentativo)* sufijoo] .o. ortografia;
```

Ejercicio 3 [36 puntos]

a) Dé una definición de constituyente sintáctico.

Un constituyente sintáctico se puede definir como una palabra o secuencia de palabras que funciona en conjunto como una unidad dentro de la estructura jerárquica de una oración.

b) Sea el conjunto de oraciones siguiente:

- *Festejemos*
- *Festejemos en el tren*
- *El ángel exterminador es una película de Buñuel*
- *Los suinos gruñen*
- *El tren de Roma a Viena partió tardíamente*
- *El sobre está sobre la mesa de luz*

i) Modele este micro universo textual utilizando GLC.

Gramática

O → GV |
GN GV
GV → V |
V GN |
V GP |
V Adv
GN → N |
Det N |
Det N GP |
Det N Adj |
Det N GP GP |
NomProp
GP → Prep GN

Léxico

V → festejemos | es | gruñen |
partió | está
Adj → exterminador
Prep → de | a | sobre | en
Det → el | una | los
N → ángel | película | suinos |
tren | sobre | mesa | luz
NomProp → Buñuel | Roma | Viena
Adv → tardíamente

Otras gramáticas son viables. Por ejemplo, "El ángel exterminador" puede manejarse como nombre propio. Se podría también manejar "mesa de luz" como un sustantivo único, aunque es más discutible.

ii) Indique al menos dos problemas que surgen al utilizar GLC para modelar lenguaje. Ejemplifique con la GLC construida en la parte anterior.

Es costoso controlar concordancia. Con la gramática anterior puede generarse, por ejemplo, la oración "el película gruñen". También hay problemas con la subcategorización. Se podría generar, por ejemplo "los suinos gruñen Buñuel".

c) Sea la GLC siguiente:

Gramática

O → GV | GN GV
GV → V GN | V GN GP
GN → Det N | Det N GP
GP → Prep GN

Léxico

N → vino | dueño | casa
Det → el | la | los
V → vino | dice
Prep → de | a

i) Aplique el algoritmo de Earley a la siguiente entrada: *vino el dueño de la casa*

chart[0]			
1	$\gamma \rightarrow \bullet$ O	[0,0]	Predecir
2	O $\rightarrow \bullet$ GV	[0,0]	Predecir
3	O $\rightarrow \bullet$ GN GV	[0,0]	Predecir
4	GV $\rightarrow \bullet$ V GN	[0,0]	Predecir
5	GV $\rightarrow \bullet$ V GN GP	[0,0]	Predecir
6	GN $\rightarrow \bullet$ Det N	[0,0]	Predecir
7	GN $\rightarrow \bullet$ Det N GP	[0,0]	Predecir
chart[1]			
8	V $\rightarrow$ vino $\bullet$	[0,1]	Buscar
9	GV $\rightarrow$ V $\bullet$ GN	[0,1]	Completar (4,8)
10	GV $\rightarrow$ V $\bullet$ GN GP	[0,1]	Completar (5,8)
11	GN $\rightarrow \bullet$ Det N	[1,1]	Predecir

12	GN → ● Det N GP	[1,1]	Predecir
chart[2]			
13	Det → el ●	[1,2]	Buscar
14	GN → Det ● N	[1,2]	Completar (11,13)
15	GN → Det ● N GP	[1,2]	Completar (12,13)
chart[3]			
16	N → dueño ●	[2,3]	Buscar
17	GN → Det N ●	[1,3]	Completar (14,16)
18	GN → Det N ● GP	[1,3]	Completar (15,16)
19	GP → ● Prep GN	[3,3]	Predecir
20	GV → V GN ●	[0,3]	Completar (9,17)
21	GV → V GN ● GP	[0,3]	Completar (10,17)
22	O → GV ●	[0,3]	Completar (2,20)
chart[4]			
23	Prep → de ●	[3,4]	Buscar
24	GP → Prep ● GN	[3,4]	Completar (19,23)
25	GN → ● Det N	[4,4]	Predecir
26	GN → ● Det N GP	[4,4]	Predecir
chart[5]			
27	Det → la ●	[4,5]	Predecir
28	GN → Det ● N	[4,5]	Completar (25,27)
29	GN → Det ● N GP	[4,5]	Completar (26,27)
chart[6]			
30	N → casa ●	[5,6]	Predecir
31	GN → Det N ●	[4,6]	Completar (28,30)
32	GN → Det N ● GP	[4,6]	Completar (29,30)
33	GP → Prep GN ●	[3,6]	Completar (24,31)
34	GN → Det N GP ●	[1,6]	Completar (18,33)
35	GV → V GN GP ●	[0,6]	Completar (21,33)
36	GV → V GN ●	[0,6]	Completar (9,34)
37	GV → V GN ● GP	[0,6]	Completar (10,34)
38	O → GV ●	[0,6]	Completar (2,35)
39	O → GV ●	[0,6]	Completar (2,36)

ii) Escriba el árbol o los árboles de análisis sintáctico correspondiente(s) al punto anterior.

Hay dos árboles posibles para la entrada, que corresponden a las filas 38 y 39 del chart anterior. Sus análisis son los siguientes:

[o [GV [V *vino*] [GN [Det *el*] [N *dueño*]] [GP [Prep *de*] [GN [Det *la*] [N *casa*]]]]]

[o [GV [V *vino*] [GN [Det *el*] [N *dueño*] [GP [Prep *de*] [GN [Det *la*] [N *casa*]]]]]]

Ejercicio 4 [6 puntos]

a) Indique 2 problemas que aparecen al representar en Lógica de Primer Orden enunciados en lenguaje natural.

b) Describa brevemente la aplicación de los métodos de aprendizaje automático supervisado y no supervisado para WSD. Mencione ventajas y desventajas de cada método.

Comentar brevemente los métodos mencionados durante el curso. Las ventajas y desventajas se encuentran resumidas en la sección 19.1 del libro del curso.

Ejercicio 5 [30 puntos]

Sea la siguiente gramática aumentada con anotaciones semánticas:

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)]
nominal → nominal adj	
nominal → nominal pp	
gv → v	[gv.sem = v.sem]
gv → v gn	[gv.sem = v.sem (gn.sem)]
pp → prep gn	
v → lee	[v.sem = λx λy ∃e leer (e) ∧ lee(e,y) ∧ leído(e,x)]
prep → de	[prep.sem = λx λy de(y,x)]
prep → en	[prep.sem = λx λy en(y,x)]

a) Complete la gramática anterior con las 3 anotaciones semánticas que faltan. Estas anotaciones deben cubrir los casos en los que los adjetivos tienen una semántica intersectiva y las preposiciones una interpretación relacional (tal como aparece en las reglas para "de" y "en").

Se agregan las partes subrayadas en 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)]
nominal → nominal adj	[nominal.sem = λx nominal.sem (x) ∧ isa (x, adj.sem)]
nominal → nominal pp	[nominal.sem = λx nominal.sem (x) ∧ pp.sem(x)]
gv → v	[gv.sem = v.sem]
gv → v gn	[gv.sem = v.sem (gn.sem)]
pp → prep gn	[pp.sem = prep.sem(gn.sem)]
v → lee	[v.sem = λx λy ∃e leer (e) ∧ lee(e,y) ∧ leído(e,x)]
prep → de	[prep.sem = λx λy de(y,x)]
prep → en	[prep.sem = λx λy en(y,x)]
npropio → Juan	[Juan]
npropio → Sartre	[Sartre]

n → autor [autor]
n → libro [libro]
adj →interesante [interesante]
adj → joven [joven]

b) Utilizando las reglas anteriores (completadas con las entradas léxicas necesarias) realice las derivaciones para las representaciones semánticas de :

i. *Juan lee un libro interesante de Sartre.*

n → libro [libro]
libro
nominal → n [λx isa(x,libro)]
libro
nominal → nominal adj [λx isa(x,libro) ∧ isa(x,interesante)]
libro interesante
pp → prep gn [λy de(y,Sartre)]
de Sartre
nominal → nominal pp [λx isa (x,libro) ∧ isa(x,interesante) ∧ de(x,Sartre)]
libro interesante de Sartre
gn → det nominal [< ∃ x isa (x,libro) ∧ isa(x,interesante) ∧ de(x,Sartre) >]
un libro interesante de Sartre
gv → v gn [λy ∃e leer (e) ∧ lee(e,y) ∧
leído(e,<∃ x isa(x,libro)∧isa(x,interesante) ∧ de(x,Sartre) >)]
lee un libro interesante de Sartre
o → gn gv [∃e leer (e) ∧ lee(e,Juan) ∧
leído(e,<∃ x isa(x,libro)∧isa(x,interesante) ∧ de(x,Sartre) >)]
Juan lee un libro interesante de Sartre

ii. *Juan lee un libro de un autor joven.*

n → libro [autor]
autor
nominal → n [λx isa(x,autor)]
autor
nominal → nominal adj [λx isa(x,autor) ∧ isa(x,joven)]
autor joven
gn → det nominal [< ∃ x isa (x,autor) ∧ isa(x,joven) >]
un autor joven
pp → prep gn [λy de(y, < ∃ x isa (x,autor) ∧ isa(x,joven) >)]
de un autor joven
nominal → nominal pp [λx isa(x,libro) ∧de(x,<∃ x isa(x,autor)∧isa(x,joven) >)]
libro de un autor joven
gn → det nominal [< ∃ y isa (y,libro) ∧
de(y,<∃ x isa(x,autor)∧isa(x,joven) >)>]
un libro de un autor joven
gv → v gn [λz ∃e leer (e) ∧ lee(e,z) ∧
leído(e, < ∃ y isa (y,libro) ∧
de(y,<∃ x isa(x,autor)∧isa(x,joven) >)>)]
lee un libro de un autor joven
o → gn gv [∃e leer (e) ∧ lee(e,Juan) ∧

leído(e, < ∃ y isa (y,libro) ∧
de(y,<∃ x isa(x,autor)∧isa(x,joven) >)>)]

Juan lee un libro de un autor joven

c) Exprese en lógica de primer orden las expresiones en *quasi logical form* obtenidas en el paso anterior.

i.

∃e leer (e) ∧ lee(e,Juan) ∧
leído(e,<∃ x isa(x,libro)∧isa(x,interesante) ∧ de(x,Sartre) >)
⇒ ∃e leer (e) ∧ lee(e,Juan) ∧
∃x leído(e,x) ∧ isa(x,libro)∧isa(x,interesante) ∧ de(x,Sartre)
⇒ ∃e ∃x leer (e) ∧ lee(e,Juan) ∧
leído(e,x) ∧ isa(x,libro)∧isa(x,interesante) ∧ de(x,Sartre)

ii.

∃e leer (e) ∧ lee(e,Juan) ∧ leído(e, < ∃ y isa (y,libro) ∧
de(y,<∃ x isa(x,autor)∧isa(x,joven) >)>)
⇒ ∃e leer (e) ∧ lee(e,Juan) ∧∃ y leído(e, y) ∧ isa (y,libro) ∧
de(y,<∃ x isa(x,autor)∧isa(x,joven) >)
⇒ ∃e leer (e) ∧ lee(e,Juan) ∧∃ y leído(e, y) ∧ isa (y,libro) ∧
∃x de(y,x) ∧isa(x,autor)∧isa(x,joven)
⇒ ∃e ∃y leer (e) ∧ lee(e,Juan) ∧ leído(e,y) ∧ isa (y,libro) ∧
de(y,x) ∧isa(x,autor)∧isa(x,joven)

d) Discuta una posible extensión de la gramática anterior para que cubra oraciones como "*Juan lee un libro en el parque*".

Esta extensión se puede hacer de modo similar a la de incorporar grupos preposicionales a los grupos nominales.

La regla en este caso sería :

gv → gv pp [gv.sem = ?]

El problema que se presenta es que la relación asociada a la preposición ("en" en este caso) debe tomar como uno de sus argumentos la variable de evento asociada al verbo ("lee" en este caso), pero está variable no está accesible (está ligada por el cuantificador existencial).

Una opción es definir un operador especial que nos permita acceder a la variable de evento. Otra opción puede ser no ligar existencialemente las variables de event, sino con operadores lambda.

Ejercicio 6 [6 puntos]

Defina brevemente las medidas *Precision*, *Recall*, *E* y F. Explique para que se usan. La medida *E* lleva un parámetro que tiene un cometido especial. ¿Cuál es y para qué se utiliza?

Las medidas de Precision y Recall, son usadas en muchas tareas de procesamiento de lenguaje natural. En particular, en el área de la Recuperación de Información sirven para medir cuan bueno es la efectividad de un sistema.
Precisión: capacidad de mostrar fundamentalmente documentos relevantes
*Recuperación: capacidad de encontrar **todos** los documentos relevantes.*

Son valores entre 0 y 1, y cuanto más próximas a 1, mejor es la evaluación de la técnica empleada. Típicamente, a mayor recall menor precision.

$$Precision = \frac{|documentos\ \ recuperados\ relevantes|}{|documentos\ \ recuperados|}$$

$$Recall\ = \frac{|documentos\ \ recuperados\ \ relevantes|}{|documentos\ \ relevantes|}$$

En cuanto a las medidas E y F, las mismas se definen en base a Precision *y* Recall*, son variantes de éstas. En particular, la medida* ***E*** *es variante de medida* ***F****. Permite poner mayor enfásis en precisión o en recuperación según el valor del parámetro*

β, el cual controla el balance entre P y R

$$F = \frac{2PR}{P+R} \qquad\qquad E = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

β = 1: Pesos iguales (En este caso ***E*** = ***F***).
β > 1: Mayor peso a precisión
β < 1: Mayor peso a recuperación