Grau en Intel.ligència Artificial

Morfología

Análisis morfológico

Detectores y correctores ortográficos

Processament del Llenguatge Humà
3 - Morfologia

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Índice

1 Morfología
- Motivación
- Definiciones
- Tipos de morfologías

2 Análisis morfológico
- Autómata de estados finitos
- Transductores de estados finitos

3 Detectores y correctores ortográficos

Índice

1 Morfología
- Motivación
- Definiciones
- Tipos de morfologías

2 Análisis morfológico
- Autómata de estados finitos
- Transductores de estados finitos

3 Detectores y correctores ortográficos

Motivation

Morfología
Motivación
Análisis morfológico
Detectores y correctores ortográficos

Necesaria para muchas herramientas y aplicaciones de PLN. Ejemplos:

- RI

  'Normally, houses in the Pyrenees are made of stone.'
  'A typical pyrenean house has litle windows.'

Motivation

Morfología
Motivación
Análisis morfológico
Detectores y correctores ortográficos

Necesaria para muchas herramientas y aplicaciones de PLN. Ejemplos:

- RI

  'Normally, houses in the Pyrenees are made of stone.'
  'A typical pyrenean house has litle windows.'

- Detectores de errores ortográficos

  'This could be an alterantive remedy'

Motivation

Morfología
Motivación
Análisis morfológico
Detectores y correctores ortográficos

Necesaria para muchas herramientas y aplicaciones de PLN. Ejemplos:

- RI

  'Normally, houses in the Pyrenees are made of stone.'
  'A typical pyrenean house has litle windows.'

- Detectores de errores ortográficos

  'This could be an alterantive remedy'

- Analizadores sintácticos

  'Children are very intelligent'
  'Children is very intelligent'

Índice

1 Morfología
- Motivación
- Definiciones
- Tipos de morfologías

2 Análisis morfológico
- Autómata de estados finitos
- Transductores de estados finitos

3 Detectores y correctores ortográficos

Definición de morfología

- Estudio de la estructura de las palabras
  - Fonología: combinación de fonemas
  - Ortografía: combinación de grafemas
  - Morfología: combinación de morfemas

Morfología
Definiciones
Análisis morfológico
Detectores y correctores ortográficos

Definición de morfología

Morfología
Definiciones
Análisis morfológico
Detectores y correctores ortográficos

- Estudio de la estructura de las palabras
  - Fonología: combinación de fonemas
  - Ortografía: combinación de grafemas
  - Morfología: combinación de morfemas
- Tipos de morfemas:
  - Raíz: (p.e., 'work', 'of','mak'[e])
  - Afijos: ocurren en combinación con otros morfemas (p.e., '-s','in-','-able')
    - Prefijos: in + frequent
    - Sufijos: work + s
    - Infijos: [Árabe] ktb + CuCuC $\rightarrow$ kutub [libros]
    - circunfijos: en+light+en

Definición de morfología

Morfología
Definiciones
Análisis morfológico
Detectores y correctores ortográficos

- Estudio de la estructura de las palabras
  - Fonología: combinación de fonemas
  - Ortografía: combinación de grafemas
  - Morfología: combinación de morfemas
- Tipos de morfemas:
  - Raíz: (p.e., 'work', 'of','mak'[e])
  - Afijos: ocurren en combinación con otros morfemas (p.e., '-s','in-','-able')
    - Prefijos: in + frequent
    - Sufijos: work + s
    - Infijos: [Árabe] ktb + CuCuC $\rightarrow$ kutub [libros]
    - circunfijos: en+light+en
- Clasificación morfo-sintáctica (*part of speech* -**PoS**- *tags*): Nombre, Verbo, Adjetivo, Adverbio, Preposición, ...

Índice

1 Morfología
- Motivación
- Definiciones
- Tipos de morfologías

2 Análisis morfológico
- Autómata de estados finitos
- Transductores de estados finitos

3 Detectores y correctores ortográficos

Tipos de morfologías

- Morfología concatenativa: concatenación de prefijos y sufijos a la raíz. Frecuente en lenguas indo-europeas.

Morfología
Tipos de morfologías
Análisis morfológico
Detectores y correctores ortográficos

Tipos de morfologías

Morfología
Tipos de morfologías
Análisis morfológico
Detectores y correctores ortográficos

- Morfología concatenativa: concatenación de prefijos y sufijos a la raíz. Frecuente en lenguas indo-europeas.
  - Morfología flexiva: *raíz* $\rightarrow$ *formas de la misma palabra*.

    P.e.: work $\rightarrow$ worked, working, works

Tipos de morfologías

Morfología
Tipos de morfologías
Análisis morfológico
Detectores y correctores ortográficos

- Morfología concatenativa: concatenación de prefijos y sufijos a la raíz. Frecuente en lenguas indo-europeas.
  - Morfología flexiva: *raíz* $\rightarrow$ *formas de la misma palabra*.

    P.e.: work $\rightarrow$ worked, working, works
  - Morfología derivativa: *raíz* $\rightarrow$ *palabras diferentes.*

    P.e.: frequent $\rightarrow$ infrequent

Tipos de morfologías

Morfología
Tipos de morfologías
Análisis morfológico
Detectores y correctores ortográficos

- Morfología concatenativa: concatenación de prefijos y sufijos a la raíz. Frecuente en lenguas indo-europeas.
  - Morfología flexiva: *raíz* $\rightarrow$ *formas de la misma palabra*.

    P.e.: work $\rightarrow$ worked, working, works
  - Morfología derivativa: *raíz* $\rightarrow$ *palabras diferentes.*

    P.e.: frequent $\rightarrow$ infrequent
  - Morfología compositiva: *N palabras* $\rightarrow$ *nueva palabra*

    P.e.: fire $+$ man $\rightarrow$ fireman

Tipos de morfologías

Morfología
Tipos de morfologías
Análisis morfológico
Detectores y correctores ortográficos

- Morfología concatenativa: concatenación de prefijos y sufijos a la raíz. Frecuente en lenguas indo-europeas.
  - Morfología flexiva: *raíz* $\rightarrow$ *formas de la misma palabra*.

    P.e.: work $\rightarrow$ worked, working, works
  - Morfología derivativa: *raíz* $\rightarrow$ *palabras diferentes.*

    P.e.: frequent $\rightarrow$ infrequent
  - Morfología compositiva: *N palabras* $\rightarrow$ *nueva palabra*

    P.e.: fire + man $\rightarrow$ fireman
- Morfología no concatenativa: otros mecanismos (infijos). Frecuente en lenguas semitas.
  - Ex: Morfología raíz-patrón

    P.e.: [Árabe] ktb + CaCaCa $\rightarrow$ kataba [él escribió]

Morfología flexiva

Los morfemas flexivos proporcionan información que depende de la PoS y el idioma

- Nombres (N):
  - Género: niñ-o (M), niñ-a (F)
  - Número: [Italiano] italian-o (SG), italian-i (PL)
  - Caso: [Alemán] die Rolle des Mann-es (Genitivo)
- Verbos (V):
  - Tiempo: want-ed (PASADO), will want (sin marca para futuro)
  - Modo: com-er (INDICATIVO), com-ed (IMPERATIVO)
  - Aspecto: want-ed (PERFECTIVO), I am waiting (sin marca para imperfectivo)
  - Voz: [Sueco] servera-s (PASIVA) [es servido]
- Adjetivos (A):
  - Género: blanc-o (M), blanc-a (F)
  - Número: blanco (SG), blanco-s (PL)
  - Comparativos: cheap-er, more similar (no todos los adjetivos)

Morfología
Tipos de morfologías
Análisis morfológico
Detectores y correctores ortográficos

Morphología derivativa

Los morfemas derivativos pueden cambiar la PoS y el significado de la palabra

- $N \rightarrow N$
  P.e.: corrección $\rightarrow$ correccion-al , camión $\rightarrow$ camion-ero
- Adjetivación: $V \rightarrow A$ or $N \rightarrow A$
  P.e.: react $\rightarrow$ react-ive, employ $\rightarrow$ employ-able
  medicine $\rightarrow$ medicin-al, use $\rightarrow$ use-ful
- Nominalización: $V \rightarrow N$ or $A \rightarrow N$
  P.e.: watch $\rightarrow$ watch-er, react $\rightarrow$ react-ion
  useful $\rightarrow$ useful-ness
- Adverbialización: $A \rightarrow Adv$
  P.e.: común $\rightarrow$ comun-mente
- Negativización:
  P.e.: frequent $\rightarrow$ in-frequent, do $\rightarrow$ un-do

Índice

1 Morfología
- Motivación
- Definiciones
- Tipos de morfologías

2 Análisis morfológico
- Autómata de estados finitos
- Transductores de estados finitos

3 Detectores y correctores ortográficos

Objetivo del análisis morfológico

Morfología

Análisis morfológico

Detectores y correctores ortográficos

- Reconocimiento léxico

  ¿Pertenece la palabra *w* al idioma *L*?

- Análisis morfológico

  ¿Qué información morfológica se asocia a la palabra *w* según el idioma *L*?

  P.e.: *palabra POS+Gen+Num+Persona+Caso+Tiempo+... LEMA (raíz)*
  men Noun+M+PL man
  cambiará Verbo+3+PL+F+IND cambiar

Recursos requeridos para el análisis morfológico

- Listas de lemas (raíces) regulares (Reg) (ambigüedades)
  P.e.: Reg_V: walk
  Reg_N: cat, fox, walk
- Listas de lemas (raíces) irregulares (Irreg) (ambigüedades)
  Ex: Irreg_pres_V: sing . . . Irreg_past_V: sang sung
  Irreg_sg_N: mouse . . . Irreg_pl_N: mice mouse
- Lista de prefijos y sufijos
  Ex: Inflexivos: -s, -ing
  Derivativos: -able, -ly, in-, un-
- Morfotáctica: reglas generales para la combinación de morfemas
  Ex: Reg_N + s $\rightarrow$ PL
  Reg_V + ing $\rightarrow$ Gerund
- Reglas de ortografía: reglas para la combinación de letras
  Ex: E-insertion: -(z,x,s,sh,ch)^s $\rightarrow$ -(z,x,s,sh,ch)es
  Consonant-doubling: -l^ing $\rightarrow$ -lling

Tipos de procesadores morfológicos

- Basado en diccionarios: lista de formas con su información morfológica
  - Ex: (write VPrI write, writes VPrI3S write, wrote VPsI write, . . .)
  - \+ eficiencia
  - \+ pueden generarse automáticamente a partir de recursos
  - \- idiomas con morfología compleja (p.e., castellano, catalán, alemán, finlandés, euskera, ...)

Tipos de procesadores morfológicos

- Basado en diccionarios: lista de formas con su información morfológica
  - Ex: (write VPrI write, writes VPrI3S write, wrote VPsI write, . . .)
  - \+ eficiencia
  - \+ pueden generarse automáticamente a partir de recursos
  - \- idiomas con morfología compleja (p.e., castellano, catalán, alemán, finlandés, euskera, ...)
- Basado en automatas de estados finitos (finite state automata -FSAs-)
  - \- solo para el reconocimiento léxico

Tipos de procesadores morfológicos

- Basado en diccionarios: lista de formas con su información morfológica
  - Ex: (write VPrI write, writes VPrI3S write, wrote VPsI write, . . .)
  - \+ eficiencia
  - \+ pueden generarse automáticamente a partir de recursos
  - \- idiomas con morfología compleja (p.e., castellano, catalán, alemán, finlandés, euskera, ...)
- Basado en automatas de estados finitos (finite state automata -FSAs-)
  - \- solo para el reconocimiento léxico
- Basados en transductores de estados finitos (finite state tranducers -FSTs-)
  - \+ útil para el análisis morfológico

Índice

1 Morfología
- Motivación
- Definiciones
- Tipos de morfologías

2 Análisis morfológico
- Autómata de estados finitos
- Transductores de estados finitos

3 Detectores y correctores ortográficos

Autómata de estados finitos (FSA)

Morfología

Análisis morfológico

Autómata de estados finitos

Detectores y correctores ortográficos

Un FSA representa la función que define la perticencia de una secuencia de símbolos, w, a un lenguaje regular L.
$M_L : w \rightarrow \{True, False\}$

$M =< Q, \Sigma, q_0, F, \sigma >$

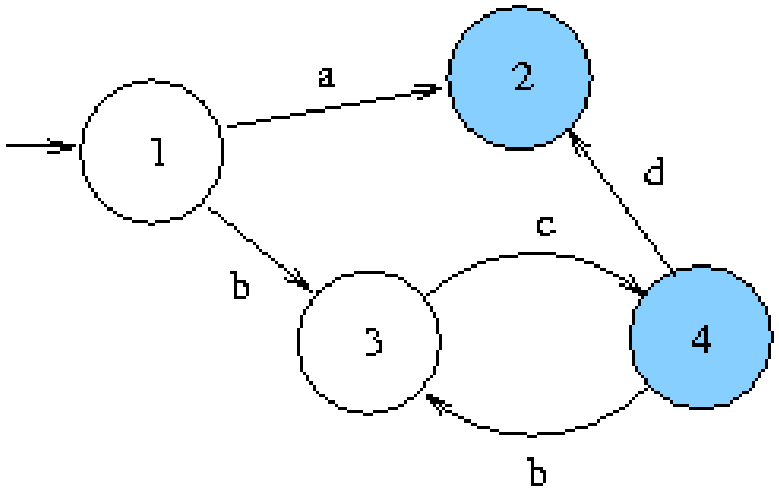
$Q = \{q_0, \ldots, q_n\}$ conjunto finito de estados

$\Sigma = \{s_0, \ldots, s_k\}$ conjunto finito de símbolos

$q_0 \in Q$ estado inicial

$F \subset Q$ conjunto de estados finales

$\sigma : Q x \Sigma \rightarrow [Q \vee 2^Q]$ función de transición [determinista $\vee$ no det.]



$a\|(bc)+d\{0,1\}$
a
bc
bcd
bcbcd
...

FSAs para reconocimiento léxico

Morfología

Análisis morfológico

Autómata de estados finitos

Detectores y correctores ortográficos

Un FSA puede ser la unión/concatenación de diferentes FSAs:

- FSAs generados a partir de reglas morfotácticas
- FSAs generados a partir de reglas ortográficas
- FSAs generados a partir de reglas derivativas
- FSAs generados a partir de reglas compositivas

FSAs para reconocimiento léxico

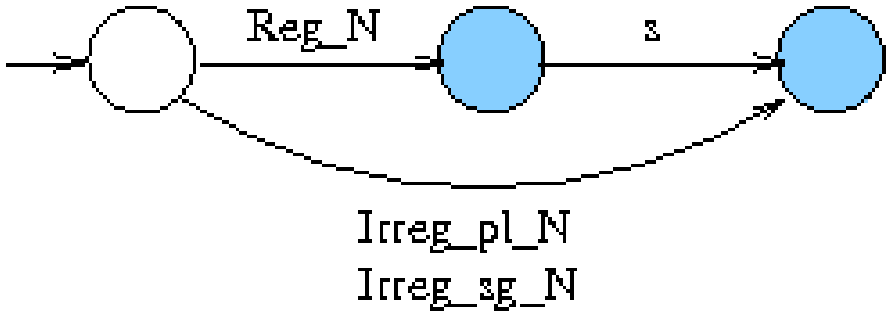
Morfología

Análisis morfológico

Autómata de estados finitos

Detectores y correctores ortográficos

P.e.: FSA generado a partir de reglas flexivas (flexión del número en nombres del inglés)



Ejemplos de listas de raices

Reg_N	Irreg_sg_N	Irreg_pl_N
dog	mouse	mice
fox	foot	feet
tax		
donkey		

FSAs para reconocimiento léxico

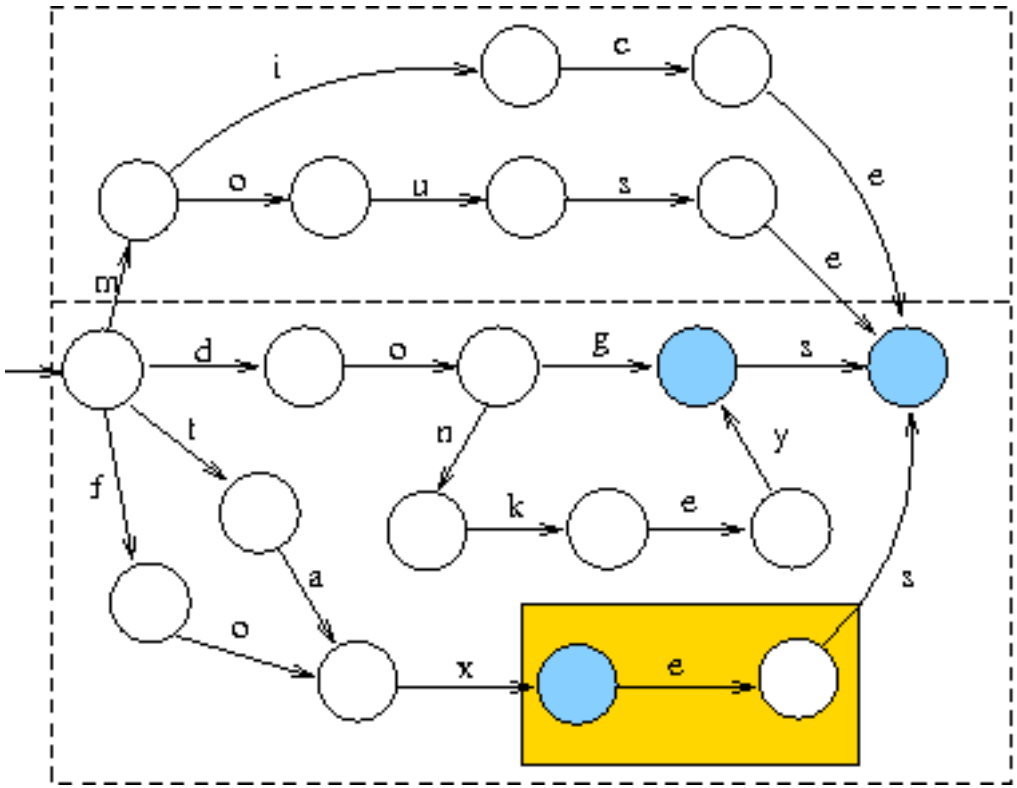
Morfología

Análisis morfológico

Autómata de estados finitos

Detectores y correctores ortográficos

P.e.: FSA para la inflexión del número en nombres comunes del inglés



Morphotactics: List Irreg_N

Morphotatics: noun + s = PL
over list Reg_N

SHOULD CORRECT WITH:

Spelling rule:
[s,x,z,sh,ch]^s=[s,x,z,sh,ch]es
over list Reg_N

FSAs para reconocimiento léxico

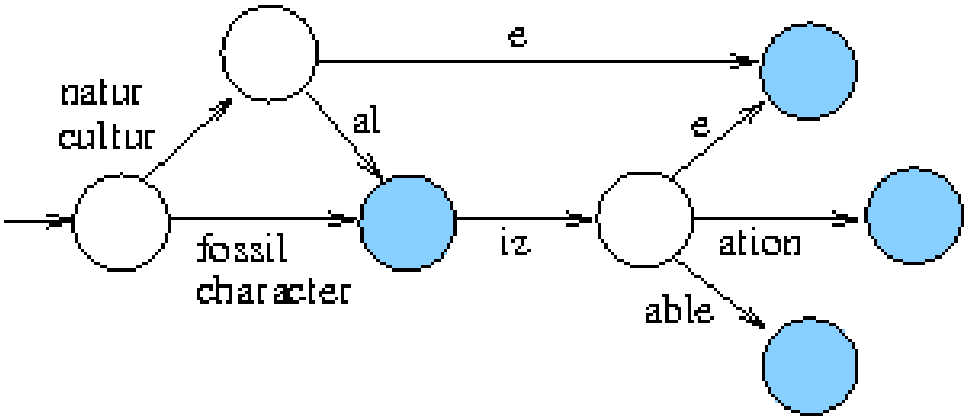
Morfología

Análisis morfológico

Autómata de estados finitos

Detectores y correctores ortográficos

P.e.: FSA generado a partir de reglas derivativas



No son tan productivas como las reglas flexivas: 'jail', 'window', ... ?

FSAs para reconocimiento léxico

Morfología

Análisis morfológico

Autómata de estados finitos

Detectores y correctores ortográficos

- FSAs pueden reconocer o generar las palabras de un idioma
- FSAs no pueden producir el análisis morfológico de una palabra

Input: forma de una palabra (forma superficial)	Output: análisis (forma léxica)
dog	dog+N+SG
dogs	dog+N+PL
(forma)	(lema+Atributos)

- Se requiere una técnica más sofisticada: FSTs

Índice

1 Morfología
- Motivación
- Definiciones
- Tipos de morfologías

2 Análisis morfológico
- Autómata de estados finitos
- Transductores de estados finitos

3 Detectores y correctores ortográficos

Transductores de estados finitos (FSTs)

Morfología

Análisis morfológico

Transductores de estados finitos

Detectores y correctores ortográficos

Un FST representa una relación entre dos lenguajes regulares $T : L_1 \rightarrow L_2$

$T =< Q, \Sigma, \Delta, q_0, F, \sigma, \delta >$

$Q = \{q_0, \ldots, q_n\}$ conjunto finito de estados

$\Sigma = \{s_0, \ldots, s_k\}$ conjunto finito de símbolos input

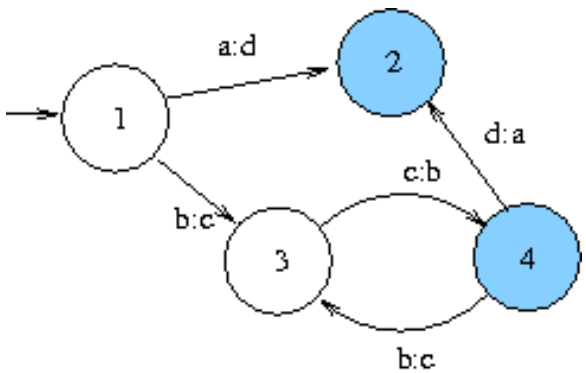
$\Delta = \{t_0, \ldots, t_m\}$ conjunto finito de símbolos output

$q_0 \in Q$ estado inicial

$F \subset Q$ conjunto de estados finales

$\sigma : Q x \Sigma \rightarrow 2^Q$ función de transición

$\delta : Q x \Sigma \rightarrow \Delta$ función de emisión



$a\|(bc)+d\{0,1\}$	$d\|(cb)+a\{0,1\}$
a	d
bc	cb
bcd	cba
bcbc	cbcb
bcbcd	cbcba
...	

Transductores de estados finitos (FSTs)

Morfología

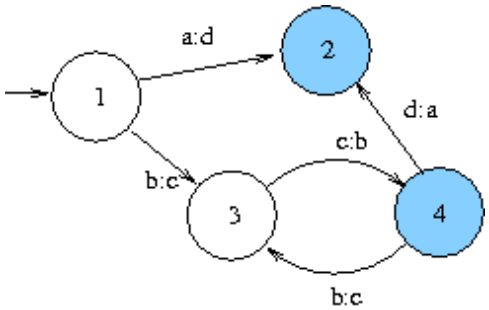
Análisis morfológico

Transductores de estados finitos

Detectores y correctores ortográficos

Propiedades relevantes:

- Inversion: $T : L_1 \rightarrow L_2 \Longrightarrow T^{-1} : L_2 \rightarrow L_1$



T : b:c $\Longrightarrow$ c $\rightarrow$ b $\Longrightarrow$ Ex: cbcb $\rightarrow$ bcbc
T^{-1}: b:c $\Longrightarrow$ b $\rightarrow$ c $\Longrightarrow$ Ex: bcbc $\rightarrow$ cbcb

- Composición: $T_a : L_1 \rightarrow L_2 \wedge T_b : L_2 \rightarrow L_3 \Longrightarrow T_a \circ T_b : L_1 \rightarrow L_3$
- x:x $\equiv$ x
- Símbolo vacio : $\epsilon \in \Sigma,\ \epsilon \in \Delta$

FSTs para el análisis morfológico

Morfología
Análisis morfológico
Transductores de estados finitos
Detectores y correctores ortográficos

Queremos un FST que represente la relación entre

- Forma superficial:

  $L_1 = \{w | w \text{ es una forma de una palabra}\}$

- Forma léxica:

  $L_2 = \{< l, F > | l \text{ es un lema} \wedge F \text{ sus atributos morfológicos}\}$

y así obtener un analizador morfológico

- P.e.: dogs $\rightarrow$ dog+N+PL
  P.e.: dog $\rightarrow$ dog+N+SG

Invirtiendo ese FST obtenemos un generador de formas

- P.e.: dog+N+PL $\rightarrow$ dogs
  P.e.: dog+N+SG $\rightarrow$ dog

FSTs para el análisis morfológico

Construcción en dos niveles:

1. T_{lex}: FST que compute la morfología flexiva y derivativa

   P.e.: Reg_N^s# $\rightarrow$ Reg_N+N+PL.
   dog^s# $\rightarrow$ dog+N+PL, fox^s# $\rightarrow$ fox+N+PL

2. T^{i}_{inter}: FSTs que computen las reglas de regularización ortográfica

   P.e.: -{z,x,s,sh,ch}es $\rightarrow$ -{z,x,s,sh,ch}^s#

Morfología
Análisis morfológico
Transductores de estados finitos
Detectores y correctores ortográficos

FSTs para el análisis morfológico

Morfología
Análisis morfológico
Transductores de estados finitos
Detectores y correctores ortográficos

Construcción en dos niveles:

1. T_{lex}: FST que compute la morfología flexiva y derivativa

   P.e.: Reg_N^s# → Reg_N+N+PL.
   dog^s# → dog+N+PL, fox^s# → fox+N+PL

2. T^i_{inter}: FSTs que computen las reglas de regularización ortográfica

   P.e.: -{z,x,s,sh,ch}es → -{z,x,s,sh,ch}^s#

Procesamiento a dos niveles:

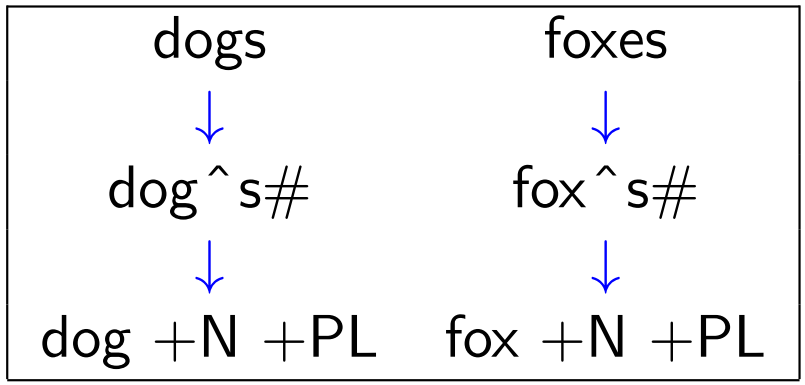
nivel superficial

$T^1_{inter}, \ldots, T^k_{inter}$

nivel intermedio

T_{lex}

nivel léxico

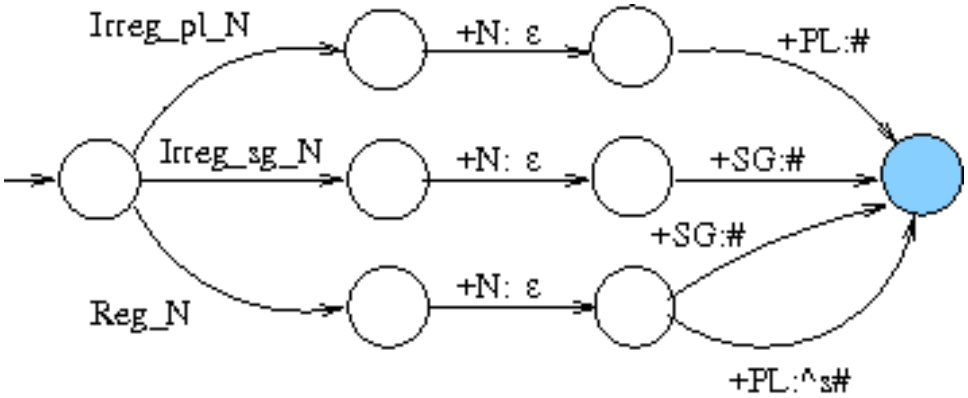


FSTs para el análisis morfológico

1. T_{lex}: FST que compute la morfología flexiva
   P.e.: FST para la flexión del número en nombres del inglés

Morfología
Análisis morfológico
Transductores de estados finitos
Detectores y correctores ortográficos

T_{num_nouns}



Ejemplos de listas de raices

Reg_N	Irreg_sg_N	Irreg_pl_N
dog	mouse	m o:i u:ϵ s:c e
fox	foot	f o:e o:e t
tax		
donkey		

FSTs para el análisis morfológico

Morfología

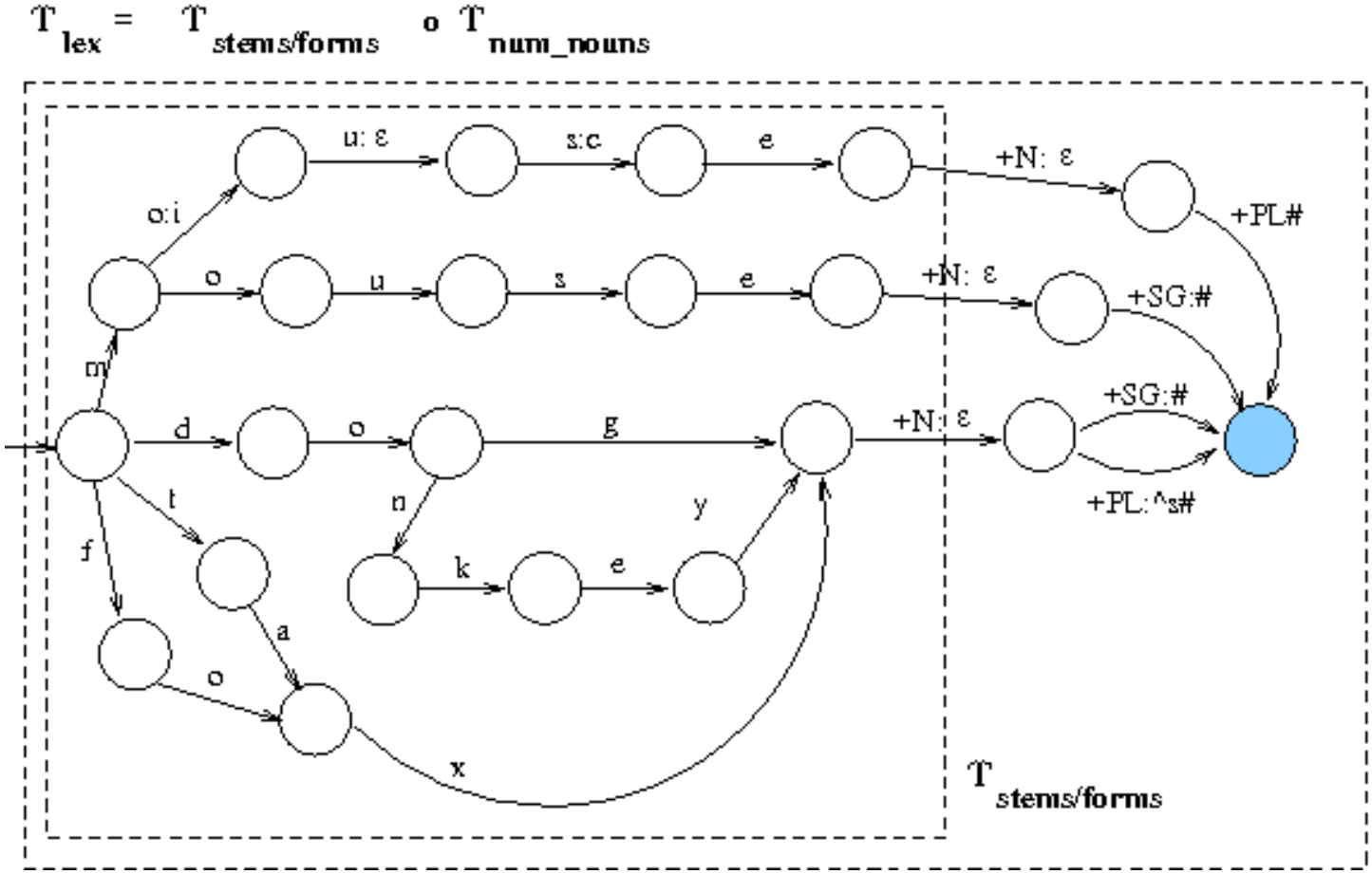
Análisis morfológico

Transductores de estados finitos

Detectores y correctores ortográficos

1 T_{lex}: FST que compute la morfología flexiva

P.e.: FST para la flexión del número en nombres del inglés



fox^s# → fox+N+PL !! (requiere regla de regularización ortográfica)

FSTs para el análisis morfológico

Morfología

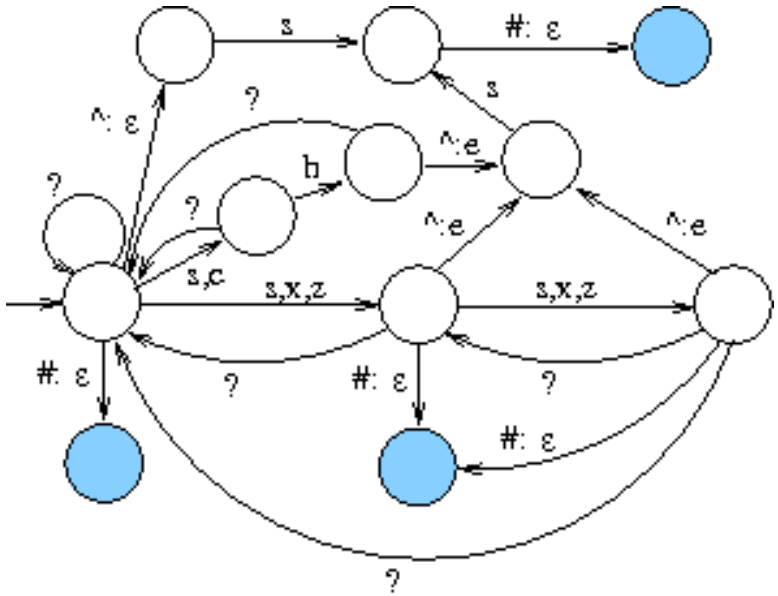
Análisis morfológico

Transductores de estados finitos

Detectores y correctores ortográficos

2 T^{i}_{inter}: FSTs que computan reglas de regularización ortográfica

P.e.: FST para la regla E-insertion



'?': otho símbolo
casos de e-insertion
foxes → fox^s#
bosses → boss^s#
flashes → flash^s#
...
casos regulares
dogs → dog^s#
...

FSTs para el análisis morfológico

Morfología

Análisis morfológico

Transductores de estados finitos

Detectores y correctores ortográficos

2 T^{i}_{inter}: FSTs que computan reglas de regularización ortográfica

Consonant doubling: duplica la útima consonante antes de *-ing/-ed* en palabras agudas de 2 sílabas acabadas con patrón CVC

P.e: control → controlling

E-deletion: elimina la *-e* muda antes de *-ing/-ed*

P.e.: remove → removed

E-insertion: *-e* added after ending *-s,-z,-x,-ch,-sh*, before *-s*

EX: flash → flashes

Y-replacement: *-y* changes to *-ie* before *-s* or to *-i* before *-ed*

EX: cry → cries, cried

K-insertion: verbs ending with *1-vowel+c* add *-k* before *-ed*

EX: panic → panicked

...

Ejercicio

Morfología

Análisis morfológico

Transductores de estados finitos

Detectores y correctores ortográficos

- Generar un FST para la flexión de los verbos *sing* and *work*
- Añadir la flexión del verbo *make* al FST previo

Índice

1 Morfología
- Motivación
- Definiciones
- Tipos de morfologías

2 Análisis morfológico
- Autómata de estados finitos
- Transductores de estados finitos

3 Detectores y correctores ortográficos

Detectores de errores ortográficos

Morfología

Análisis morfológico

Detectores y correctores ortográficos

- **Objetivo**: dado un texto, reconocer las palabras que no pertenecen al lenguaje del texto
- **Posible método**:

FSA_L (o FST_L)
$S = Tokenizer(text)$ (secuencia de palabras)
for each $x \in S$
if $FSA_L(x)$ then print("x")
else print("$**x**$")

Correctores ortográficos

Morfología

Análisis morfológico

Detectores y correctores ortográficos

- **Objetivo**: dado un texto, corregir las palabras que no pertenecen al lenguaje del texto
- **Tarea inicial**: dada una palabra, proporcionar una lista de posibles palabras correctas
- **Posible método**:

$D = \{y_i : y_i \in L\}$ generado via FST_L
$S = Tokenizer(text)$ (secuencia de palabras)

for each $x \in S$
if $x \in D$ then print(x)
else
$D' = \{y \in D : |length(x) - length(y)| \leqslant \gamma\}$
$C = \emptyset$
for each $y \in D'$
$d = distance(x, y)$
if $(d \leqslant \delta)$ then $C = C + \{< y, d >\}$
print_Nbest_candidates(C,N)

($\delta = 2$ y $\gamma = 2$ parecen suficientes para texto estándar)

Correctores ortográficos

- Distancia de edición: mínimo número de inserciones, eliminaciones, intercambios de caracteres para obtener y desde x

Morfología

Análisis morfológico

Detectores y correctores ortográficos

Correctores ortográficos

Morfología

Análisis morfológico

Detectores y correctores ortográficos

- Distancia de edición: mínimo número de inserciones, eliminaciones, intercambios de caracteres para obtener y desde x
- Distancia de edición ponderada: mínimo coste de inserciones, eliminaciones, intercambios de caracteres para obtener y desde x
  - Insertar o eliminar un caracter $= 1$
  - Intercambiar dos caracteres $= s(c1, c2)$ (distancia de Manhattan en un teclado)
  - coste_total(x,y) $= E_{m,n}$:
    Computar la matriz de coste, E, de dimensiones mXn (longitudes de x e y) usando programación dinámica

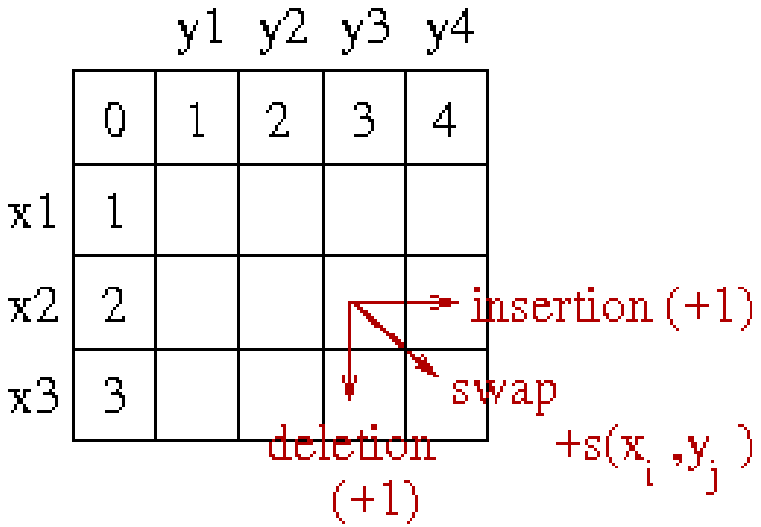
Correctores ortográficos

Morfología

Análisis morfológico

Detectores y correctores ortográficos

Computación de la matriz de coste



$$E_{i,j} = min(Cost_{del}, Cost_{ins}, Cost_{swap})$$

$$\begin{cases} Cost_{del} = E_{i-1,j} + 1 \\ Cost_{ins} = E_{i,j-1} + 1 \\ Cost_{swap} = E_{i-1,j-1} + s(x_i, y_j) \end{cases}$$

$s(x_i, y_j)$	a	b	c	d	e
a	0				
b	0.5	0			
c	0.3	0.3	0		
d	0.2	0.2	0.1	0	
e	0.3	0.4	0.2	0.1	0

$s(x_i, y_j)$ normalizada a 1.0

Ejercicio

Morfología

Análisis morfológico

Detectores y correctores ortográficos

- Computar la distancia de edición ponderada entre 'dom' and 'come'