Concatenación y Desconcatenación: Operaciones Fundamentales de la Lingüística
Notas

Fernando Galindo Soria 13 jun 96

-Concatenación (definiciones básicas)
. alfabeto/ elementos palabras
. concatenación
. cadena
. concatenación entre alfabetos
. A^2
. transitividad de la concatenación
. A^n
. longitud de la cadena
. cadena vacía
. A^+
. A^* =| vocabulario V
. Lenguaje subconjunto o igual de V
. Si alfa pertenece a L, alfa es una oración
. concatenación entre cadenas
. subcadena
. unidad léxica
- Ejemplos de concatenación
el perro ladra
Concatenación de mapas de bits: .juegos por computadora, navegantes en realidad virtual
[FIGURA ONDA DE SONIDO] el perro ladra

- Desconcatenación
. dada una cadena separarla en subcadenas
. puede desconcatenarse de múltiples formas
Ejemplo
el perro ladra
el perro ladra
el perro ladra
el p err o la dra
. separada por blancos
En general el blanco no existe
.Buenos Aires
.corrector ortográfico
de cuatro en cuatro
etc.

=>
Algunos tipos de desconcatenación deseable
por elementos comunes
por unidades léxicas (cadena con significado propio)
Ejemplo

Dibuja un Árbol y una nube

	PALABRA	SIGNIFICADO	ETIQUETA
=>	Dibuja	ignora	i
	un	ignora	i
X	árbol	rutina árbol	a
	y	ignora	i
	una	ignora	i
=>	nube	rutina nube	n

i i a i i n

=>

Y

***** *****

Desconcatena palabra por palabra (separada por blancos) y la sustituye por su significado
Concatena significados

-Concatenador de mapas de bits (para juegos, etc.)
-Concatenador de voces
hola /\/_____/_
como /__/\/_
estas __/\/\/\/\

hola como estas => X => /\/_____/_/__/\/___/\/\/\/\

-Complejidad del problema de la desconcatenación
Dibuja un árbol (desconcatena palabras separadas por blanco)
el perro ladra (reconocedor de caracteres, delimita (desconcatena) mapa de bits (palabras))

separadas por el fondo (blanco)

reconocedor de voz delimita (desconcatena) palabras separadas por silencios

señal donde aparentemente no existe un blanco (un motor funcionando, electrocardiograma, etc.) (en algunos casos se puede encontrar un “blanco” pero no siempre)

Señales que integran a varias (las señales se solapan)

Algunas herramientas de desconcatenación
-separación por blanco
-delimitación de la cadena
a) Explicita: cursor variable
Voy a Buenos Aires con Juan Manuel
(se marca la cadena: de donde empieza a donde termina y se corta)

b) Delimitador Variable (Jesús M. Olivares) explícitamente se delimita la cadena y se corta

Herramienta básica: cortar y pegar
delimitación implícita de la cadena
-por blancos
-Por el fondo. el sistema busca una cadena y la delimita por el fondo
- por silencios, el sistema corta cadenas separadas por silencios

Por patrones conocidos (matcheo)
Buenos Aires A
Juan Manuel B

El sistema busca las subcadenas dentro de cadenas ya conocidas y corta por esos puntos.

Estoy en Buenos Aires llamando a Juan Manuel en Bogota

=>

Estoy en A llamando a B en Bogota

donde A y B son puntos de corte

-Método de Cuitlahuac Cantú
frecuencia de elementos
frecuencia de subcadenas
Método de Cuitlahuac

Resta de patrones conocidos
señal
cloro
=>
señal sin cloro

Introducción a desconcatenación de cadenas
a) Cadenas en ASCII
Primera opción: aceptar el blanco
marcar la cadena (buscar siempre de mayor a menor)
b) letras separadas
cursor variable de Jesús M. Olivares

c) letra manuscrita
d) palabras o frases cerradas

-Desconcatenación de señales en general (trucos)
.Cursor Variable
.Rejilla (Cuitlahuac)
n x m

movible permanentemente (izq./ der./ arriba/ abajo/ inclinándose/ etc.)
(aleatoriamente en espacio controlado
libremente)
=> siempre acumula los valores

-Problema del desconcatenamiento
.Desconcatenar cadena
el perro corre
a) por letras e l p e r r o c o r r e
b) por palabras separadas por blanco el perro corre
los componentes de una cadena se llaman subcadenas
c) En general el blanco no es un separador

En Buenos Aires se estudia Música
-Concepto de Unidad léxica
Una unidad léxica es una cadena de caracteres (o una señal o una subcadena) con significado propio
Dado que se tiene un conjunto de Unidades Léxicas, yo puedo representar cada unidad léxica por una etiqueta y dada una cadena puedo sustituir las unidades léxicas por sus etiquetas.
=>
oración canónica

Ejemplo:

Dadas las siguientes oraciones
O-> el perro blanco
O-> el perro negro
O-> el gato blanco
=>
puedo proponer las siguientes unidades léxicas y sustituciones

A-> el
B-> perro
C-> blanco
D->negro
E->gato
O->ABC
O->ABD
O->AEB

En lingüística se acostumbra poner primero las oraciones
O->ABC
O->ABD
O->AEB
A-> el
B-> perro

C-> blanco
D->negro
E->gato

sistema de reescritura con unidades léxicas

ABC
oración canónica

En general representar el lenguaje como cadenas de unidades léxicas es mas compacto que las oraciones explicitas

Ejercicio

Oraciones => X cadenas de unidades léxicas => Y => Oraciones

Hacer un programa que me genera el sistema de reescritura de unidades léxicas

Hacer un programa que a partir del sistema de reescritura de unidades léxicas genera oraciones
Los síntomas de un sistema experto se pueden ver como un conjunto de vectores

Sintoma1 peso1 Sintoma2 peso2 ... Sintomam pesom
S1 P1 S2 P2 S3 P3 . . . Sm Pm -> Diagnostico
y si se grafican dan una curva (firma) característica

(El numero del síntoma se puede ver como un ángulo y el peso como una magnitud ¿o al revez?)
y cada grafica de un color representa un diagnostico

Se puede normalizar dividiendo entre el numero de acumulados antes de graficar.