Universidad Nacional de San Agustín de Arequipa
Escuela Profesional de Ciencia de la Computación
Docentes:
M.Sc. Carlos Eduardo Atencio Torres

Estimación de parámetros (probabilidades de las reglas)

Existen 2 modos

– **Usar un corpus de árboles (como el Penn Treebank)**

• La probabilidad de una regla se aproxima por la frecuencia relativa de su utilización en el corpus.

La probabilidad de una regla se aproxima por la frecuencia relativa de su utilización en el corpus

$$P(A \rightarrow \alpha) = C(A \rightarrow \alpha) / \Sigma\beta\ C(A \rightarrow \beta) = C(A \rightarrow \alpha) / C(A)$$

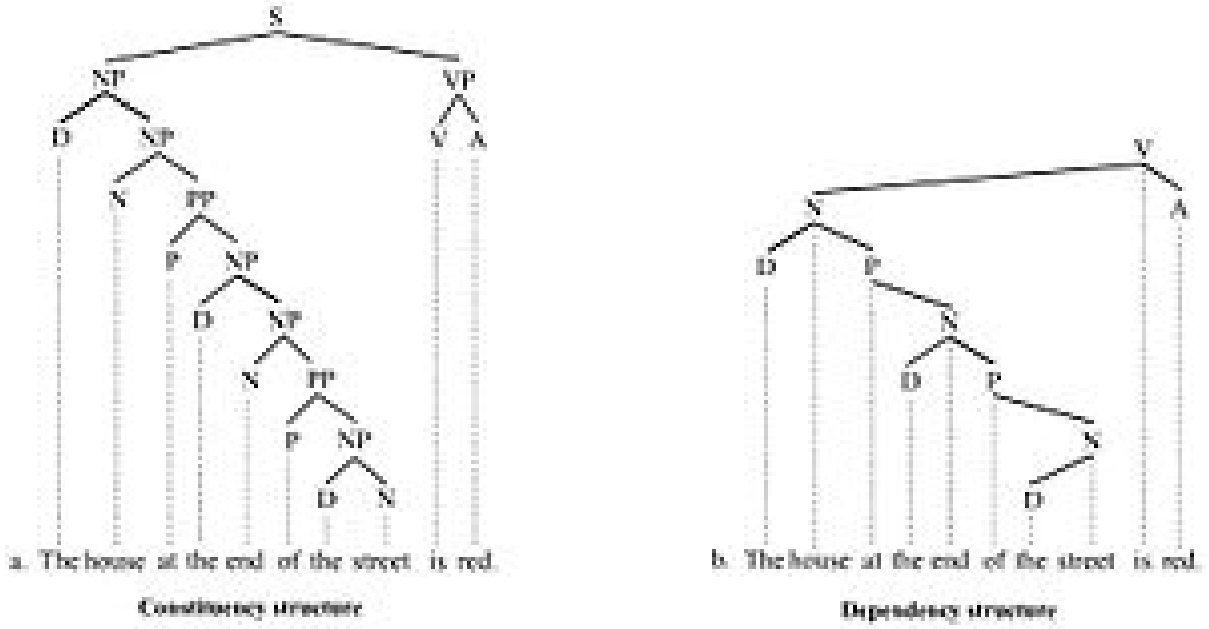
C(A → α) : cantidad de veces que se usa la regla A → α en el corpus de árboles

C(A) : cantidad de veces que aparece el símbolo A en el corpus

Corpus anotados -Treebanks

• Los treebanks son corpus en los que cada oración está asociada a un árbol sintáctico.

• Se pueden crear: – Directamente, anotando “a mano”. – Con parsing automático y posterior corrección manual.



Penn Treebank

• El Penn Treebank es un corpus anotado ampliamente usado (inglés), mantenido por el LDC (Linguistic Data Consortium).

• Contiene árboles de análisis con información sintáctica y algo de información semántica – una base de datos de árboles lingüísticos.

Categorías gramaticales (tagset)

15 categorías distintas (corpus Brown)

• Incluyen variantes en número para sustantivos

• NN sustantivo singular cat

• NNS sustantivo plural cats

En inglés no hay variación en género (solo pronombres) y ni los adjetivos ni los determinantes varían en número.

• Incluyen variantes en forma, persona y tiempo para verbos

• VB forma base eat

• VBD pasado ate

• VG gerundio eating

• VBN participio eaten

• VBP presente, no 3era persona eat

• VBZ presente, 3era persona eats

Category	Genre (Code)	# of texts	Total Tokens	%
INFORMATIVE	Learned (J)	80	160,000	16.0%
INFORMATIVE	Belles Lettres, Biography, Memoirs, etc (G)	75	150,000	15.0%
INFORMATIVE	Popular Lore (F)	48	96,000	9.6%
INFORMATIVE	Press: Reportage (A)	44	88,000	8.8%
INFORMATIVE	Skills and Hobbies (E)	36	72,000	7.2%
INFORMATIVE	Miscellaneous (H)	30	60,000	6.0%
IMAGINATIVE	General Fiction (K)	29	58,000	5.8%
IMAGINATIVE	Adventure and Western Fiction (N)	29	58,000	5.8%
IMAGINATIVE	Romance and Love Story (P)	29	58,000	5.8%
INFORMATIVE	Press: Editorial (B)	27	54,000	5.4%
IMAGINATIVE	Mystery and Detective Fiction (L)	24	48,000	4.8%
INFORMATIVE	Press: Reviews (theatre, books, music, dance) (C)	17	34,000	3.4%
INFORMATIVE	Religion (D)	17	34,000	3.4%
IMAGINATIVE	Humor (R)	9	18,000	1.8%
IMAGINATIVE	Science Fiction (M)	6	12,000	1.2%
	TOTAL	500	1,000,000	100.0%

Penn Treebank Tagset

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

A partir de un corpus se puede definir una gramática, en donde se trata de tomar todos los sub-árboles locales en el conjunto de árboles del corpus. Usando el Pen podemos inferir una gramática como por ejemplo:

```
...
(NP
      (NP (DT a) (NN round))
      (PP (IN of)
       (NP
        (NP (JJ similar) (NNS increases))
        (PP (IN by)
         (NP (JJ other) (NNS lenders)))
        (PP (IN against)
         (NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
...
```

Reglas "inferidas" de los árboles

NP → NP PP

NP → DT NN

NP → JJ NNS

NP → NP PP PP

NP → NNP JJ NN NNS (estructuras “chatas”)
PP → IN NP

Hallando las probabilidades

Tenemos una GLC, con reglas como: NP → NP PP

• Extraemos estas reglas del treebank.

• Estimamos las probabilidades de las reglas :

p(NP → NP PP) = cantidad(NP → NP PP) cantidad(NP)

• Obtenemos una GLCP.

– Calcularlas sobre un corpus no anotado

• Se comienza con reglas equiprobables.

• Se recalculan las probabilidades según resultados del parsing del paso anterior.

• Se itera hasta converger. (Algoritmo inside-outside)