Gramáticas Formales para el Lenguaje Natural

Análisis sintáctico probabilístico

Análisis sintáctico probabilístico

- Insuficiencias de GLC
- GLC probabilísticas (GLCP)
  - Definición
  - Probabilidad de un árbol y de una tira
  - Usos
  - Métodos de parsing
- De dónde vienen las probabilidades?
  - Corpus anotados con árboles
  - Penn TreeBank
  - Gramática inferida de un corpus, ejemplo del Penn
- Problemas con GLCP, algunas soluciones
  - La aplicación de una regla no es independiente del lugar en el árbol
  - El léxico incide en la forma preferida del árbol sintáctico
  - Soluciones: anotación con padres, subcategorización?
- Medidas de evaluación
  - *Precision / recall / F*
  - Paréntesis cruzados

Problemas con el enfoque de GLC

- Concordancia - multiplicidad de reglas
  - GN → Det Nom (x 4 + “el agua”)

- Subcategorización
  - *Juan quiere salir. // * Juan comió + Vinf*
  - *Juan dijo la verdad. // * Juan dijo la mesa.*
  - *Presencié el incendio. // * Presencié la silla.*

  No todas las combinaciones de categorías gramaticales son posibles, las reglas deben preverlas todas.

- **Dificultad en tener una gramática de cobertura completa**

- **Múltiples análisis – ambigüedad**

  Juan vio al hombre en el parque con el telescopio

  ¿Cuál es el análisis correcto?

Problemas con el enfoque de GLC

Para sistemas de parsing (90's)

- Los parsers asociados a GLC tienen una cobertura pobre (40% o más de oraciones sin análisis).
- Incluso oraciones simples tienen muchos análisis.
  - Los parsers no tienen método de selección entre los distintos análisis.
- Una propuesta de solución a este problema:
  - GLC extendidas con un mecanismo de puntaje.
  - Parsing que selecciona el análisis mejor punteado.
  - Ha dado resultado (con diversas modificaciones y extensiones).

Ranqueo de árboles

Se propone un enfoque tal que:

1- No se escribe una gramática, sino que se escriben árboles sintácticos para un conjunto amplio de oraciones (un corpus anotado). Con suficientes casos, podremos aspirar a una buena cobertura. Los ejemplos anotados definen de modo implícito la gramática.

2- Se utiliza la similitud con los árboles del corpus como medida de bondad (*gold standard*) para el análisis de una oración nueva. Utilizaremos la probabilidad del árbol como medida. El paradigma es GLCP: gramáticas libres de contexto probabilísticas.

Gramáticas libres del contexto probabilísticas (GLCP)

G = (T, N, S, R, P)

- T conjunto de terminales
- N conjunto de no terminales
- S símbolo inicial
- R conjunto de reglas de la forma $X \rightarrow \gamma$,
  - » X es un no terminal
  - » γ es una secuencia de terminales y no terminales
- P(r) da la probabilidad de cada regla r

Se cumple:

$$\forall X \in N, \sum_{X \rightarrow \gamma \in R} P(X \rightarrow \gamma) = 1$$

Gramáticas libres del contexto probabilísticas (GLCP)

La gramática G genera un modelo de lenguaje L (conjunto de tiras con sus probabilidades)

$$\sum_{\gamma \in T^*} P(\gamma) = 1$$

Probabilidad de un árbol

- P(*t*) – La probabilidad de un árbol es el producto de las probabilidades de las reglas usadas para generarlo.

- Se asume independencia en la aplicación de las distintas reglas.

- Notar que P(t,O) = P(t) P(O|t) = P(t).1

Ejemplo (adaptado de Manning)

S	→ NP VP	1.0	
VP	→ V NP	0.7	
VP	→ VP PP	0.3	
PP	→ P NP	1.0	
P	→ *con*	1.0	
V	→ *comió*	1.0	
NP	→ NP PP	0.4	
NP	→ *Juan*	0.1	
NP	→ *María*	0.18	
NP	→ *arroz*	0.04	
NP	→ *cuchara*	0.18	
NP	→ *queso*	0.1	

NP = Noun Phrase

VP = Verbal Phrase

PP = Prepositional Phrase

Probabilidades de árboles y tiras

w_{15} = *Juan comió arroz con cuchara*

Probabilidades de árboles y tiras

w_{15} = *Juan comió arroz con cuchara*

t1= [[Juan] [comió [arroz [con cuchara]]]

t2= [[Juan] [comió [arroz] [con cuchara]]

Probabilidades de árboles y tiras

- w_{15} = *Juan comió arroz con cuchara*

- $P(t_1)$ = 1.0 * 0.1 * 0.7 * 1.0 * 0.4 * 0.18 * 1.0 * 1.0 * 0.18 = 0.0009072

- $P(t_2)$ = 1.0 * 0.1 * 0.3 * 0.7 * 1.0 * 0.18 * 1.0 * 1.0 * 0.18 = 0.0006804

- $P(w_{15}) = P(t_1) + P(t_2)$
  = 0.0009072 + 0.0006804
  = 0.0015876

Probabilidad de una tira

- P(w_{1n}) – La probabilidad de una tira es la suma de las probabilidades de los árboles que tienen esa tira como alcance.

- La probabilidad de una tira interesa para modelos de lenguaje (transcripción de habla, de escritura manuscrita, etc.).

Parsing probabilístico

- Producir el árbol más probable (o los k más probables).
- Los algoritmos de parsing son extensiones de los algoritmos para GLC.
- Basados en general en CKY (gramática en forma normal de Chomsky).
- Necesitamos una GLCP : ¿cómo calculamos las probabilidades de transición?

Estimación de parámetros (probabilidades de las reglas)

2 modos

- Usar un corpus de árboles
  - La probabilidad de una regla se aproxima por la frecuencia relativa de su utilización en el corpus.

- 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*)

Estimación de parámetros (probabilidades de las reglas)

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

$$P(A \rightarrow \alpha) = C(A \rightarrow \alpha) / \Sigma_{\beta} C(A \rightarrow \beta) =$$
$$C(A \rightarrow \alpha) / C(A)$$

$C(A \rightarrow \alpha)$: cantidad de veces que se usa la regla $A \rightarrow \alpha$ 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.*

- En parsing se trabajó fundamentalmente con la sección del Wall Street Journal : 1Millón de palabras (unas 40.000 oraciones) de ediciones del Wall Street Journal del período 1987-1989.

PENN TREEBANK

```
( (S
    (NP-SBJ (DT The) (NN move))
    (VP (VBD followed)
      (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))))))
      (, ,)
      (S-ADV
        (NP-SBJ (-NONE- *))
        (VP (VBG reflecting)
          (NP
            (NP (DT a) (VBG continuing) (NN decline))
            (PP-LOC (IN in)
              (NP (DT that) (NN market)))))))
    (. .)))
```

The move followed a round of similar increases by other lenders against Arizona real estate loans, reflecting a continuing decline in that market.

La decisión se tomó tras una ronda de aumentos similares de otros proveedores de fondos de préstamos de bienes raíces de Arizona, lo que refleja una disminución continua en ese mercado. (traductor de Google)

PENN TREEBANK

```
( (S
    (NP-SBJ (DT The) (NN move))
    (VP (VBD followed)
      (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))))))
      (, ,)
      (S-ADV
        (NP-SBJ (-NONE- *))
        (VP (VBG reflecting)
          (NP
            (NP (DT a) (VBG continuing) (NN decline))
            (PP-LOC (IN in)
              (NP (DT that) (NN market)))))))
    (. .)))
```

The move followed a round of similar increases by other lenders against Arizona real estate loans, reflecting a continuing decline in that market.

Categorías gramaticales (*tagset*)

- 45 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*

  6 tags bastan

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*			

Gramática a partir de corpus

- Los corpus anotados con árboles definen implícitamente una gramática.
- Se trata de tomar todos los sub-árboles locales en el conjunto de árboles del corpus.
- Si el corpus es suficientemente grande, la cobertura de la gramática será bastante buena.
- Pero se depende de los criterios utilizados para anotar con árboles las oraciones del corpus.
- En el caso del Penn, se suelen tener estructuras muy chatas y gran cantidad de reglas.

```
…
(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))))))
...
```

¿Cuáles son las reglas que podemos inferir de este 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

Aprendiendo la gramática : reglas

…

```
(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

Aprendiendo la gramática : probabilidades

- Tenemos una GLC, con reglas como:

$$NP \rightarrow NP\ PP$$

- Extraemos estas reglas del treebank.

- Estimamos las probabilidades de las reglas :

$$p(NP \rightarrow NP\ PP) = \frac{\text{cantidad}(NP \rightarrow NP\ PP)}{\text{cantidad}(NP)}$$

- Obtenemos una GLCP.

CKY Probabilístico

Similar a CKY para GLC

- Gramática en Chomsky Normal Form (CNF)

  Reglas:

  A -> B C o A -> w

- Representamos la entrada con índices entre palabras:

  $_0$ Tomo $_1$ un $_2$ vuelo $_3$ a $_4$ París $_5$

- Pos[i,j,A] , (n+1)x(n+1)xV de la matriz contiene probabilidad de A entre i y j

CKY Probabilístico

Similar a CKY GLC

- Gramática en Chomsky Normal Form (CNF)

  Reglas:

  - A -> B C o A -> w
- Representamos la entrada con índices entre palabras:

  $_0$ Tomo $_1$ un $_2$ vuelo $_3$ a $_4$ París $_5$

- Pos[i,j,A] , (n+1)x(n+1)xV de la matriz contiene probabilidad de A entre i y j

Matriz CKY de 3 dimensiones

Algoritmo CKY Probabilístico

function PROBABILISTIC-CKY(*words*,*grammar*) **returns** most probable parse
and its probability

for $j \leftarrow$ **from** 1 **to** LENGTH(*words*) **do**
  for all $\{ A \mid A \rightarrow words[j] \in grammar \}$
    $table[j-1, j, A] \leftarrow P(A \rightarrow words[j])$
  for $i \leftarrow$ **from** $j-2$ **downto** 0 **do**
    for $k \leftarrow i+1$ **to** $j-1$ **do**
      for all $\{ A \mid A \rightarrow BC \in grammar,$
        and $table[i,k,B] > 0$ **and** $table[k,j,C] > 0 \}$
        if $(table[i,j,A] < P(A \rightarrow BC) \times table[i,k,B] \times table[k,j,C])$ **then**
          $table[i,j,A] \leftarrow P(A \rightarrow BC) \times table[i,k,B] \times table[k,j,C]$
          $back[i,j,A] \leftarrow \{k,B,C\}$
  return BUILD_TREE($back[1, \text{LENGTH}(words), S]$), $table[1, \text{LENGTH}(words), S]$

Jurafsky and Martin – Speech and Language Processing

Algoritmo CKY Probabilístico

```
function PROBABILISTIC-CKY(words,grammar) returns most probable parse
                                                and its probability
  for j ← from 1 to LENGTH(words) do
    for all { A | A → words[j] ∈ grammar}
      table[j − 1, j, A] ← P(A → words[j])
    for i ← from j − 2 downto 0 do
      for k ← i + 1 to j − 1 do
        for all { A | A → BC ∈ grammar,
                      and table[i,k,B] > 0 and table[k, j,C] > 0 }
          if (table[i,j,A] < P(A → BC) × table[i,k,B] × table[k,j,C]) then
            table[i,j,A] ← P(A → BC) × table[i,k,B] × table[k,j,C]
            back[i,j,A] ← {k,B,C}
  return BUILD_TREE(back[1, LENGTH(words), S]), table[1, LENGTH(words), S]
```

Se recorre por columnas, en sentido ascendente.

Jurafsky and Martin – Speech and Language Processing

Algoritmo CKY Probabilístico

function PROBABILISTIC-CKY(*words*,*grammar*) **returns** most probable parse
and its probability

for $j \leftarrow$ **from** 1 **to** LENGTH(*words*) **do**
  for all $\{ A \mid A \rightarrow words[j] \in grammar \}$
    $table[j-1, j, A] \leftarrow P(A \rightarrow words[j])$
  for $i \leftarrow$ **from** $j-2$ **downto** 0 **do**
    for $k \leftarrow i+1$ **to** $j-1$ **do**
      for all $\{ A \mid A \rightarrow BC \in grammar,$
        and $table[i,k,B] > 0$ **and** $table[k,j,C] > 0 \}$
        if $(table[i,j,A] < P(A \rightarrow BC) \times table[i,k,B] \times table[k,j,C])$ **then**
          $table[i,j,A] \leftarrow P(A \rightarrow BC) \times table[i,k,B] \times table[k,j,C]$
          $back[i,j,A] \leftarrow \{k,B,C\}$
  return BUILD_TREE($back[1, \text{LENGTH}(words), S]$), $table[1, \text{LENGTH}(words), S]$

Se recorre por columnas, en sentido ascendente.

Se guarda solo el mejor resultado intermedio.

Jurafsky and Martin – Speech and Language Processing

Problemas con GLCP

Las estimaciones de probabilidades son inadecuadas porque:

- Las hipótesis de independencia en las GLC no son realistas.
- No hay intervención del léxico, no puede reflejarse p.ej. que *dar* admite CD y CI mientras que *correr* es intransitivo:
  - *Juan da un caramelo a un niño.*
  - *Juan corre rápido.*
- Los modelos probabilísticos actuales usan alguna versión extendida de GLCP, o modifican la gramática del Penn TB.

Dependencias entre reglas

- En una GLC, la expansión de un no terminal es independiente del contexto.
- En una GLCP, la probabilidad de una regla es independiente del resto del árbol.
- Pero el modo en que se expande un nodo puede depender de su ubicación en el árbol.

Dependencias entre reglas

- Ejemplo: en inglés es mucho más común un pronombre en posición sujeto que objeto.

- Según datos de un corpus:

	Pronombre	No-pronombre
Sujeto	91%	9%
Objeto	34%	66%

Dependencias entre reglas

- Pero si miramos la probabilidad total de ambas reglas en el corpus :

NP → DT NN	.28
NP → PRP	.25

- Para reflejar la influencia de la posición en el árbol sobre las probabilidades de las reglas se usa la técnica de **anotación con padres.**

Dependencias léxicas

- Falta de sensibilidad de las probabilidades respecto a las palabras.

- Las palabras juegan un rol (reglas preterminales), pero no es suficiente.

- Considerar:
  - *Entregaron los paquetes al sereno.*
  - *Comieron escalopes al marsala.*

  Hay 2 análisis en c/ejemplo, el verbo informa sobre cuál es el análisis preferido.

Dependencias léxicas

- *Entregaron los paquetes al sereno.*
- *Comieron escalopes al marsala.*

análisis 1:
GV → V NP PP probabilidad p1

análisis 2:
GV → V NP probabilidad p2
NP → NP PP probabilidad p3

GV → V NP → V NP PP probabilidad p2 * p3

Según las frecuencias en el corpus, un parser va a preferir siempre una de las 2 opciones.

La expansión de GV y de NP no depende del léxico.

Dependencias léxicas

- *Entregaron los paquetes al sereno.*
- *Comieron escalopes al marsala.*

análisis 1:
GV → V NP PP

análisis 2:
GV → V NP
NP → NP PP

entregar algo a alguien
comer algo

Las estructuras argumentales influyen en las reglas gramaticales que componen el árbol.

Coordinación

(Ejemplo Collins, categorías del Penn)

dogs in houses and cats

Reglas:
NP → NP CC NP
NP → NP PP
NP → NNS
PP → IN NP

NNS → dogs | cats | houses
CC → and
IN → in

Coordinación

(Ejemplo Collins, categorías del Penn)

dogs in houses and cats

Reglas:

NP → NP CC NP (1)

NP → NP PP (2)

NP → NNS (3)

PP → IN NP (4)

NNS → dogs | cats | houses (5) (6) (7)

CC → and (8)

IN → in (9)

2 análisis

$[[dogs\ [in\ houses]_{PP}]_{NP}\ and\ [cats]_{NP}]_{NP}$

$[[dogs]_{NP}\ [in\ [\ houses\ and\ cats]_{NP}]_{PP}]_{NP}$

Reglas idénticas, siempre la misma probabilidad

Mejoras a GLCP

- “Contextualizar” las reglas independientes del contexto.

- Lexicalizar.

Mejoras a GLCP
Partir no terminales (contextualizar)

- Partimos la categoría NP en 2 versiones (ejemplo NP sujeto y objeto)
  - NP-sujeto
  - NP-objeto

- Un modo de implementar esta idea es hacer **anotación con el padre** (Johnson, 1998).

- Se renombran los nodos especializándolos con la categoría del padre:
  - Un NP sujeto va a ser NP^S (el NP es hijo del nodo S)
  - Un NP objeto va a ser NP^VP (el NP es hijo de un nodo VP)

- Las derivaciones a partir de NP^S y NP^VP tienen ahora conteos diferentes (*ejemplo con pronombres*).

Mejoras a GLCP
Partir no terminales

- Se pueden hacer además especializaciones adicionales.
- Problemas:
  - Aumenta el tamaño de la gramática (cantidad de no-terminales).
  - Se reduce entonces la cantidad de datos para entrenamiento.

Mejoras a GLCP
Lexicalización

- En lugar de modificar la gramática, se modifica el modelo probabilístico.

entregar algo a alguien	| GV → V GN GP
comer algo	| GV → V GN (OJO!, el GN puede incluir GP)

- El núcleo del constituyente (la palabra) incide en la estructura sintáctica preferida (para un GV el núcleo es el verbo, para un GN el nombre, etc.).

GV(entrega) → V (entrega) GN(paquete) GP(a)

GV(come) → V(come) GN(escalopes GP(al marsala))

Mejoras a GLCP
Lexicalización

GV(entrega) → V (entrega) GN(paquete) GP(a)
GV(come) → V(come) GN(escalopes)

$$\frac{\text{cantidad(GV(entrega) → V (entrega) GN(paquete) GP(a))}}{\text{cantidad(GV(entrega))}}$$

- En el Penn hay unas 40.000 oraciones y unas 12.500 reglas, sin lexicalizar incluso hay datos muy dispersos.
- Lexicalizando no se pueden juntar estadísticas, hay demasiada dispersión.
- Se hacen nuevas hipótesis de independencia, de modo de estimar las probabilidades de una regla a partir de cantidades razonables.

Mejoras a GLCP
Lexicalización

p(GV(entrega) → V (entrega) GN(paquetes) GP(a))

se aproxima en base a los valores:

- cantidad(GV(entrega))
- cantidad(GN(paquetes) bajo un nodo GV(entrega))
- cantidad(GP(a) bajo un nodo (GV,entrega))

O sea, se condiciona en el núcleo (en este caso GV(entrega)), y se cuentan por separado los complementos (condicionados por el núcleo).

Evaluación de parsers

Varios factores a medir

- Correctitud
  - Aceptar sólo las tiras correctas.
  - Dar el análisis adecuado a las tiras correctas.
- Eficiencia
- Escalibilidad

Evaluación de parsers

Correctitud

- Equivalencia fuerte respecto al *gold standard.*
- Generalmente la comparación se hizo respecto al Penn Treebank.
- En muchos casos hay necesidad de realizar transformaciones para poder comparar.

Evaluación de parsers

Medidas de comparación

recall = constituyentes correctos algoritmo /
total constituyentes correctos referencia

precision = constituyentes correctos algoritmo/
total constituyentes algoritmo

Un constituyente es correcto si tiene el mismo alcance y etiqueta que un constituyente en la referencia (en general, la referencia es el *gold standard*).

Evaluación de parsers

Medidas de comparación

Medida F – media armónica entre *recall* (R) y *precision* (P)

$$F = 2PR/(P+R)$$

P y R tienden a crecer en sentidos distintos.

Evaluación de parsers

Medidas de comparación

Paréntesis cruzados (*crossing brackets*) :

Constituyentes para los cuales la referencia tiene parentizado ((A B) C) y el resultado de nuestro algoritmo (A (B C)).

Evaluación de parsers

Actualmente, la performance de parsers entrenados y testeados en la sección WSJ del Penn Treebank es:

- precision > 90%
- recall 90%
- paréntesis cruzados – 1% por oración

(datos no actualizados al 2017)

Referencias

- Jurafsky, D. y J. Martin. 2009. *Speech and Language Processing*. Second edition. Capítulo 14
- Manning C., H. Schutze, 1999 *Foundations of Statistical Natural Language Processing* Massachussets Institute of Technology, USA