Gramáticas Categoriales Combinatorias

Parsing

GFLN

InCo

2016

Parsing Probabilístico para CCG

1. Generación de un corpus de derivaciones a partir del Penn
2. Supertagging - categorías léxicas complejas
3. Parsing con CKY
4. Ambigüedad espúrea, criterios para preferir derivaciones
5. Aprendizaje
6. Open-CCG

Penn TreeBank, versión II

1993, Universidad de Pennsylvania
Características generales

- 1M palabras (40.000 oraciones)
- Anotado manualmente, 3 años de trabajo
- Anotación chata
- Distinción no muy clara de complementos y adjuntos
- Elementos vacíos, necesarios para dependencias.

Penn Treebank, ejemplos

```
(S (NP-SBJ Casey)
     (VP throws
         (NP the ball)))
```

Casey throws the ball

Oración simple

Se anota la función sujeto, pero no objeto directo cuando sigue directamente al verbo

Penn Treebank, ejemplos

```
(S (NP-SBJ Casey)
   (VP should
       (VP have
           (VP thrown
               (NP the ball)))))
```

Casey should have thrown the ball

Penn Treebank, ejemplos

```
(S (NP-SBJ-1 Casey)
     (VP ought
         (S (NP-SBJ *-1)
            (VP to
                (VP have
                    (VP thrown
                        (NP the ball)))))))
```

Casey ought to have thrown the ball
Control sujeto, co-indización

Penn Treebank, ejemplos

```
(S (NP-SBJ-1 The ball)
   (VP was
       (VP thrown
           (NP *-1)
           (PP by
               (NP-LGS Casey)))))
```

The ball was thrown to Casey

Voz pasiva, co-indización del sujeto con posición objeto del verbo.

Derivaciones CCG a partir del Penn Treebank

Referencia: generación de CCGbank a partir del Penn

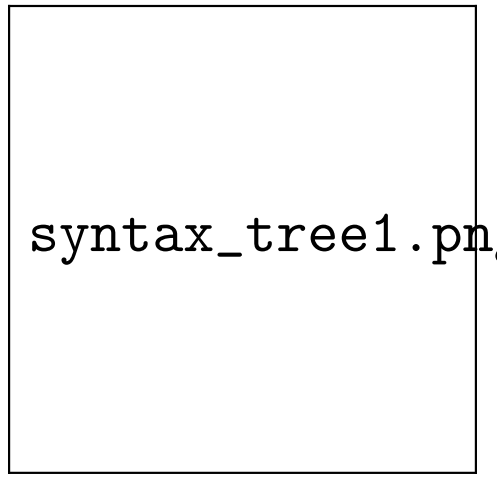
Julia Hockenmaier (2003)
Data and Models for Statistical Parsing with Combinatory Categorial Grammar, PhD thesis, University of Edinburgh

- Se transforma un árbol del Penn en una derivación CCG + relaciones de dependencia
- Algoritmo de traducción
  1. Identificar tipo del nodo (núcleo, argumento, adjunto)
  2. Binarizar el árbol
  3. Obtener categorías CCG
  4. Obtener estructura de dependencia

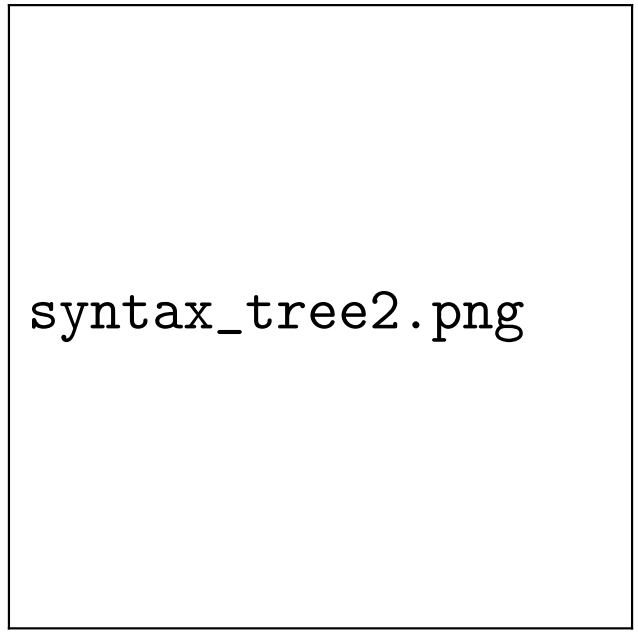
Ejemplo de traducción

Ejemplo:

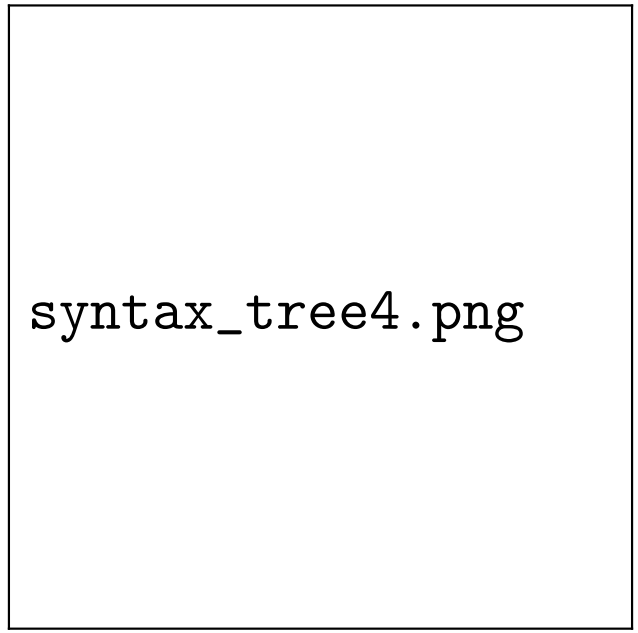
- ... just opened its doors in July



Árbol Penn con tipos agregados



árbol binarizado



se sustituyen las categorías, de modo top down

Algunos aspectos de la traducción

- Atributos en CCGbank
- VPs infinitivas, de participio y de gerundio
- Control y *raising*
- *Small clauses*
- Inversión
- Apositivas y coordinación
- Cuantificadores complejos
- Coordinación
- Parentéticas
- Extraposición de apositivas
- Expresiones multi palabra
- Adjuntos clausales

Categorías, objetos categoriales y dependencias en CCGbank

Categorías

- Cuatro categorías básicas: S, NP, N, PP
- Categorías complejas: las formadas por abstracción funcional sobre otras categorías (básicas y complejas): ej. S/NP

Categorías, objetos categoriales y dependencias en CCGbank

Categorías

- Cuatro categorías básicas: S, NP, N, PP
- Categorías complejas: las formadas por abstracción funcional sobre otras categorías (básicas y complejas): ej. S/NP

Objetos categoriales

Estructuras de datos para almacenar rasgos y atributos compartidos de las categorías, ya sean básicas o complejas.

Categorías, objetos categoriales y dependencias en CCGbank

Relaciones de dependencia

Una relación de dependencia es una 3-tupla $<< c, w >, i, < c', w' >>$ donde c es una categoría funcional con aridad $\geq i$, w una palabra y w' es la cabeza léxica del i-ésimo argumento de c.
Los argumentos de las categorías complejas se numeran de 1 a n, empezando por el más interior : (S \NP1)/NP2
Las relaciones de dependencia generadas al parsear se juntan: el parser se evalúa por estas relaciones de dependencia.

Atributos en CCGBank, oración (S)

Se proponene atributos para diferenciar subclases de la categoría básica S y de la compleja S\NP (VP).
Se usan tanto en oraciones principales como en subordinadas (o en fragmentos).

Atributos en CCGBank, oración (S)

Se proponene atributos para diferenciar subclases de la categoría básica S y de la compleja S\NP (VP).
Se usan tanto en oraciones principales como en subordinadas (o en fragmentos).

- S[dcl]: oraciones declarativas
- S[wq]: interrogativas qu- (wh-)
- S[q]: interrogativas si/no
- S[qemb]: interrogativas embebidas
- S[emb]: declarativas embebidas
- S[frg]: fragmentos (derivados de la etiqueta FRAG del Treebank)
- S[for]: claúsulas reducidas introducidas por *for*
- S[intj]: interjecciones
- S[inv]: inversión elíptica

Atributos en CCGBank, oración (S)

Estos atributos se usan para especificar información de subcategorización:

- el verbo *doubt* toma declarativas embebidas (*doubt that*) e interrogativas (*doubt whether*) como complemento

Atributos en CCGBank, oración (S)

Estos atributos se usan para especificar información de subcategorización:

- el verbo *doubt* toma declarativas embebidas (*doubt that*) e interrogativas (*doubt whether*) como complemento
- el verbo think solo toma declarativas (*think that*)

Atributos en CCGBank, oración (S)

Estos atributos se usan para especificar información de subcategorización:

- el verbo *doubt* toma declarativas embebidas (*doubt that*) e interrogativas (*doubt whether*) como complemento
- el verbo think solo toma declarativas (*think that*)
- tipo de that : S[emb]/S[decl]

Atributos de S en frase verbal (VP = S\NP)

- S[b]\NP: infinitivas sin *to*, subjunctivas e imperativas
  I suggest that he study.

Atributos de S en frase verbal (VP = S\NP)

- S[b]\NP: infinitivas sin *to*, subjunctivas e imperativas
  I suggest that he study.
- S[to]\NP: infinitivas con *to*
  She wants to run.

Atributos de S en frase verbal (VP = S\NP)

- S[b]\NP: infinitivas sin *to*, subjunctivas e imperativas
  I suggest that he study.
- S[to]\NP: infinitivas con *to*
  She wants to run.
- S[pass]\NP: participiales
  The bridge was closed for repair.

Atributos de S en frase verbal (VP = S\NP)

- S[b]\NP: infinitivas sin *to*, subjunctivas e imperativas
  I suggest that he study.
- S[to]\NP: infinitivas con *to*
  She wants to run.
- S[pass]\NP: participiales
  The bridge was closed for repair.
- S[ng]\NP: de gerundio
  He enjoys reading.

VPs infinitivas, de participio y de gerundio

- En el Penn, las frases verbales de participio, de gerundio, imperativas e infinitivas se anotan como S con sujeto nulo, que puede estar coindizado con otro elemento.

Infinitiva

```
a. (NP (NP the policy)
           (S (NP-SBJ (-NONE- *))
               (VP (TO to)
                     (VP  (VB seduce)
                            (NP socialist nations)
                            (PP-CLR into the capitalist sphe
```

VPs infinitivas, de participio y de gerundio

- En el Penn, las frases verbales de participio, de gerundio, imperativas e infinitivas se anotan como S con sujeto nulo, que puede estar coindizado con otro elemento.

Gerundio

```
b. (S (NP-SBJ-1 The banks)
        (VP (VBD stopped)
              (S (NP-SBJ (-NONE- *-1))
                   (VP (VBG promoting)
                         (NP the packages))))))
```

VPs infinitivas, de participio y de gerundio

- Los auxiliares y modales toman una frase verbal como complemento

Construcción con auxiliar

```
(S (NP-SBJ Both sides)
    (VP (VBP are)
        (VP (VBG taking)
            (NP action))))
```

Problemas en la traducción

- No está marcada la distinción complemento adjunto, hay que deducirla
- Problemas al binarizar, por ejemplo las estructuras con N N N, comunes en inglés *(goverment finances report* , quién modifica a quién)
- Proliferación de categorías para una palabra
  - pre y post modificadores
  - tipos elevados o no

CCGbank, estadísticas

- Léxico de aprox 75.000 entradas, 930.000 tokens.

CCGbank, estadísticas

- Léxico de aprox 75.000 entradas, 930.000 tokens.
- Número promedio de categorías per token: 19,2.

CCGbank, estadísticas

- Léxico de aprox 75.000 entradas, 930.000 tokens.
- Número promedio de categorías per token: 19,2.
- 1286 tipos de categorías léxicas:

CCGbank, estadísticas

- Léxico de aprox 75.000 entradas, 930.000 tokens.
- Número promedio de categorías per token: 19,2.
- 1286 tipos de categorías léxicas:
  - 439 categorías ocurren solo una vez
  - 556 ocurren 5 o más veces

CCGbank, estadísticas

Hay derivaciones para el 99.4 de las oraciones en el Penn

Palabras con mayor número de categorías:

Pal	Cats	Freq	Pal	Cats	Freq
as	130	4237	of	59	22782
is	109	6893	that	55	7951
to	98	22056	who	52	1140
than	90	1600	not	50	1288
in	79	15085	are	48	3662
-	67	2001	with	47	4214
's	67	9249	so	47	620
for	66	7912	if	47	808
at	63	4313	on	46	5112
was	61	3875	from	46	4437

Supertagging

- Para parsear una oración, se comienza por asignar una categoría a cada palabra.
- Este proceso se conoce como *supertagging*, ya que las etiquetas son estructuras sintácticas detalladas.

Supertagging

- Para parsear una oración, se comienza por asignar una categoría a cada palabra.
- Este proceso se conoce como *supertagging*, ya que las etiquetas son estructuras sintácticas detalladas.
- Se ha hablado de *almost parsing*, ya que el parser tiene menos trabajo para hacer por el tipo de etiqueta.

Dificultades del supertagging

- Más de 1000 categorías vs. unas 50 categorías en el Penn Treebank.

Dificultades del supertagging

- Más de 1000 categorías vs. unas 50 categorías en el Penn Treebank.
- Posible Baseline: para cada palabra usar el tag más frecuente

Dificultades del supertagging

- Más de 1000 categorías vs. unas 50 categorías en el Penn Treebank.
- Posible Baseline: para cada palabra usar el tag más frecuente
  - En el Penn, ese baseline da un 90% de accuracy

Dificultades del supertagging

- Más de 1000 categorías vs. unas 50 categorías en el Penn Treebank.
- Posible Baseline: para cada palabra usar el tag más frecuente
  - En el Penn, ese baseline da un 90% de accuracy
  - En CCGbank, se obtiene solo un 72%

Multi-supertagging

- Se necesita que el tagger asigne más de una categoría en algunos casos.

Multi-supertagging

- Se necesita que el tagger asigne más de una categoría en algunos casos.
- Se utiliza (Clark y Curran, 2004) un parámetro α que regula la cantidad de categorías por palabra, de modo dinámico

Multi-supertagging

- Se necesita que el tagger asigne más de una categoría en algunos casos.
- Se utiliza (Clark y Curran, 2004) un parámetro α que regula la cantidad de categorías por palabra, de modo dinámico
- Assignar categoría C si $p(\mathrm{C}|s) > \alpha.p(\mathrm{C}_{max}|s)$, C_{max} es la categoría que da valor máximo

Multi-supertagging adaptativo

- La cantidad de categorías asignada a una palabra es un factor crucial en la velocidad del tagger.

Multi-supertagging adaptativo

- La cantidad de categorías asignada a una palabra es un factor crucial en la velocidad del tagger.
- Curran & Clark proponen una estrategia "adaptativa":

Multi-supertagging adaptativo

- La cantidad de categorías asignada a una palabra es un factor crucial en la velocidad del tagger.
- Curran & Clark proponen una estrategia "adaptativa":
  - Comenzar con α alto, o sea, ambiguedad baja.

Multi-supertagging adaptativo

- La cantidad de categorías asignada a una palabra es un factor crucial en la velocidad del tagger.
- Curran & Clark proponen una estrategia "adaptativa":
  - Comenzar con α alto, o sea, ambiguedad baja.
  - Si el parser falla en encontrar un análisis, decrementar α.
  - Repetir hasta que se encuentre un análisis o reportar falla por exceso de tiempo.

El problema de la ambiguedad espúrea

- El problema de la ambiguedad espúrea

El problema de la ambiguedad espúrea

- El problema de la ambiguedad espúrea
- Multiplicidad de categorías en el léxico (por variación de orden en la oración)

El problema de la ambiguedad espúrea

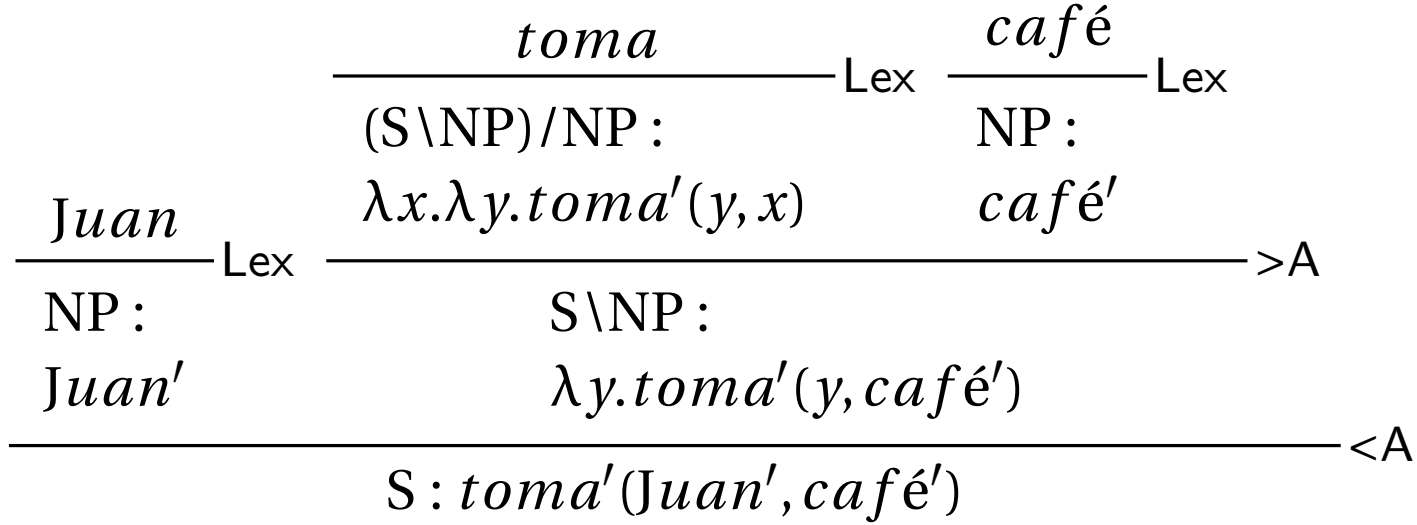
- El problema de la ambiguedad espúrea
- Multiplicidad de categorías en el léxico (por variación de orden en la oración)
- Se utiliza CKY, de modo de manejar compactada la ambigüedad

Varias derivaciones equivalentes

$$\dfrac{\dfrac{\dfrac{Juan}{\mathrm{NP}}\ \mathrm{Lex} \qquad \dfrac{\dfrac{toma}{(\mathrm{S}\backslash\mathrm{NP})/\mathrm{NP}}\ \mathrm{Lex} \qquad \dfrac{café}{\mathrm{NP}}\ \mathrm{Lex}}{\mathrm{S}\backslash\mathrm{NP}}\ >\mathrm{A}}{\mathrm{S}}\ <\mathrm{A}$$

Varias derivaciones equivalentes

con semántica



Varias derivaciones equivalentes

con semántica

$$\dfrac{\dfrac{Juan}{\text{NP}:\ Juan'}\,\text{Lex} \quad \dfrac{\dfrac{toma}{(\text{S}\backslash\text{NP})/\text{NP}:\ \lambda x.\lambda y.toma'(y,x)}\,\text{Lex} \quad \dfrac{café}{\text{NP}:\ café'}\,\text{Lex}}{\text{S}\backslash\text{NP}:\ \lambda y.toma'(y,café')}\,>\text{A}}{\text{S}:\ toma'(Juan',café')}\,<\text{A}$$

La derivación con elevación de tipos y composición (ya visto) tiene la misma semántica

Este problema se conoce como ambigüedad espúrea.

El problema de la ambiguedad espúrea

- La composición combinada con elevación de tipos genera varios análisis con la misma semántica.
- Soluciones propuestas
  - chequear la equivalencia con la interpretación semántica (puede ser exponencial)
  - restricciones durante el parsing:
    - usar elevación de tipos+composición solo cuando es imprescindible
    - usar restricciones de forma normal durante el parsing

Parsing estadístico CCG

- Se han aplicado modelos generativos (2002,2003) y condicionales (2004,2007).

Parsing estadístico CCG

- Se han aplicado modelos generativos (2002,2003) y condicionales (2004,2007).
- Resultados: 83% de medida F con DepBank como gold standard

Parsing estadístico CCG

- Se han aplicado modelos generativos (2002,2003) y condicionales (2004,2007).
- Resultados: 83% de medida F con DepBank como gold standard
- Atributos para el aprendizaje
  - standard (POS, word) se usan para el baseline, que solo considera árboles locales (4 tipos)
  - relaciones palabra-palabra inferidas de estructura predicado-argumento (dependencias locales) y de dependencias no locales
  - otros atributos léxicos: head (la palabra)
  - otros atributos estructurales : padre del padre, medidas de distancia

Referencias

- Clark, S. y JR Curran. 2004. The Importance of Supertagging for Wide-Coverage CCG Parsing. ACL 2004.
- Hockenmaier, J. 2003. Data and Models for Statistical Parsing with Combinatory Categorial Grammar, PhD thesis, University of Edinburgh
- Hockenmaier, J. y M. Steedman. 2007. CCGBank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. Computational Linguistics. Volumen 33, Nro. 3.