Inducción No Supervisada de Gramáticas de Lenguaje Natural

Franco M. Luque

Grupo de Procesamiento de Lenguaje Natural
Universidad Nacional de Córdoba & CONICET
Córdoba, Argentina

Primera Jornada de Doctorandos de Computación
FaMAF, Universidad Nacional de Córdoba
6 de diciembre de 2010

Introducción

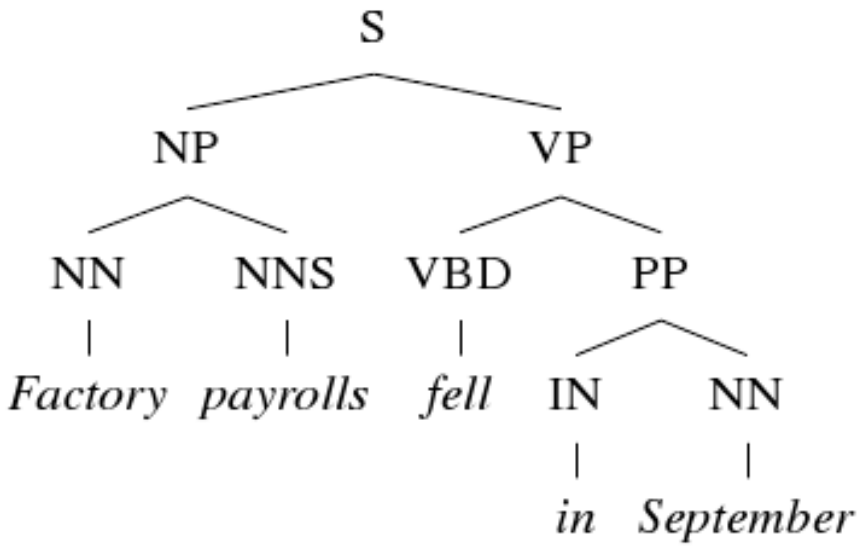
- Lic. en Ciencias de la Computación, FaMAF (2000 - 2007).
- Estudiante del Doctorado en Computación desde abril de 2007 (ahora terminando el 4to año de Doctorado).
- Becario Tipo II del Conicet hasta marzo 2012.
- Director: Gabriel Infante-Lopez.
- Tema de Trabajo: "Inducción No Supervisada de Gramáticas de Lenguaje Natural".
- Otras actividades:
  - Profesor Asistente ded. simple: Paradigmas, Algoritmos I, Ingeniería I (desde abril 2008).
  - Co-director con Gabriel de trabajo de grado sobre PLN.

Resumen de la Charla

Análisis Sintáctico (Parsing)

El Análisis Sintáctico (Parsing) trata el problema de, dada una oración de lenguaje natural, obtener su estructura sintáctica de acuerdo a cierta teoría lingüística.

- Por ejemplo: Árboles de constituyentes.



- Usualmente también se incluye una categorización de las palabras con los llamados *POS (Part Of Speech) tags*.

Parsing Supervisado

El Parsing Supervisado trata el problema de construir un parser usando un *treebank*.

- Los *treebanks* son corpus de oraciones parseadas (por tipos que saben).
- Una parte del *treebank* se utiliza para entrenar el parser.
- Otra parte se utiliza para evaluarlo.

Puede ser visto como inducir una función $f : X \rightarrow Y$ (el parser) conociendo algunos puntos $(x_1, f(x_1)), \ldots, (x_n, f(x_n))$ (el *treebank*).

Parsing No Supervisado

¿Qué podemos hacer conociendo sólo un conjunto $\{x_1, \ldots, x_n\}$ del dominio de f?

- El parser se entrena sólo con oraciones.
- La evaluación se sigue haciendo con un *treebank*.

Parsing No Supervisado

¿Qué podemos hacer conociendo sólo un conjunto $\{x_1, \ldots, x_n\}$ del dominio de f?

- El parser se entrena sólo con oraciones.
- La evaluación se sigue haciendo con un *treebank*.

Motivaciones:

- Construir parsers baratos y treebanks preliminares para idiomas que no cuentan con *treebanks*.
- Buscar argumentos para la discusión acerca de la naturaleza del aprendizaje del lenguaje en la niñez.

Parsing No Supervisado

¿Qué podemos hacer conociendo sólo un conjunto $\{x_1, \ldots, x_n\}$ del dominio de f?

- El parser se entrena sólo con oraciones.
- La evaluación se sigue haciendo con un *treebank*.

Motivaciones:

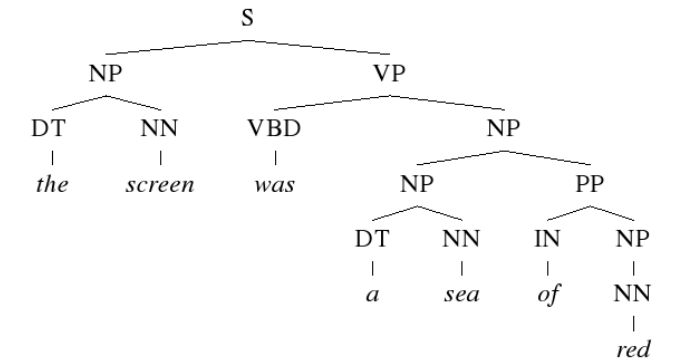
- Construir parsers baratos y treebanks preliminares para idiomas que no cuentan con *treebanks*.
- Buscar argumentos para la discusión acerca de la naturaleza del aprendizaje del lenguaje en la niñez.

Estado del arte:

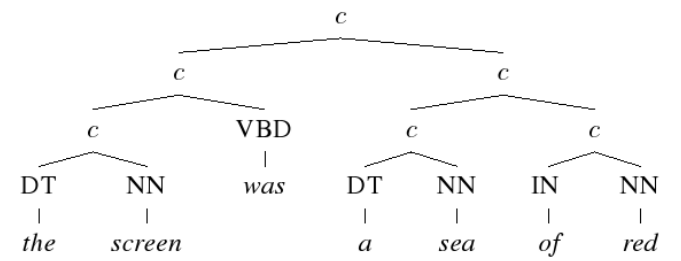
- Resultados alentadores en *unlabeled parsing* de oraciones cortas (Klein & Manning, 2005).
- Posteriores avances en parsing de dependencias (Martincito).

Evaluación de Parsers

Árbol "de oro":



Árbol propuesto:



- *Precision*: proporción de constituyentes propuestos que estan bien $(4/5)$.
- *Recall*: proporción de constituyentes correctos que fueron encontrados $(4/5)$.
- F1: Media harmónica entre *precision* y *recall*.

Inferencia Gramatical

El área de Inferencia Gramatical estudia el problema de descubrir un lenguaje formal oculto, perteneciente a una clase de lenguajes conocida, a través de cierto conjunto de evidencias:
Posibles evidencias:

- Ejemplos positivos (como en el Parsing No Supervisado),
- ejemplos negativos,
- oráculo que responde a consultas,
- otras, combinaciones, etc.

Inferencia Gramatical

El área de Inferencia Gramatical estudia el problema de descubrir un lenguaje formal oculto, perteneciente a una clase de lenguajes conocida, a través de cierto conjunto de evidencias:
Posibles evidencias:

- Ejemplos positivos (como en el Parsing No Supervisado),
- ejemplos negativos,
- oráculo que responde a consultas,
- otras, combinaciones, etc.

Una clase de lenguajes es *aprendible* si existe un algoritmo que siempre encuentra el lenguaje oculto a través de la evidencia.

Aprendibilidad

Identificabilidad en el Límite (IIL) (Gold, 1967):

- Sólo se asume que toda evidencia eventualmente aparece.
- El algoritmo eventualmente encuentra el lenguaje oculto.
- La clase de lenguajes finitos es IIL con evidencia positiva.
- Los CFLs no son IIL con evidencia positiva.

Aprendibilidad

Identificabilidad en el Límite (IIL) (Gold, 1967):

- Sólo se asume que toda evidencia eventualmente aparece.
- El algoritmo eventualmente encuentra el lenguaje oculto.
- La clase de lenguajes finitos es IIL con evidencia positiva.
- Los CFLs no son IIL con evidencia positiva.

Aprendibilidad Probablemente Aproximadamente Correcta (PAC):

- La evidencia se muestrea de acuerdo a una distrib. prob.
- A más evidencia, el algoritmo más se acerca al lenguaje oculto.
- Los CFLs probabilísticos son PAC-aprendibles con evidencia sólo positiva (Horning, 1969).

El Lenguaje Natural como Lenguaje Formal

Trata el problema de formalizar el Lenguaje Natural como una clase de lenguajes formales. Un enfoque posible:

- Alfabeto: las palabras del lenguaje (no las letras).
- Lenguaje: conjunto de oraciones sintácticamente correctas.
- Adecuación fuerte: la formalización debe asignar la estructura sintáctica correcta (los árboles) a las oraciones.

El Lenguaje Natural como Lenguaje Formal

Trata el problema de formalizar el Lenguaje Natural como una clase de lenguajes formales. Un enfoque posible:

- Alfabeto: las palabras del lenguaje (no las letras).
- Lenguaje: conjunto de oraciones sintácticamente correctas.
- Adecuación fuerte: la formalización debe asignar la estructura sintáctica correcta (los árboles) a las oraciones.

El Lenguaje Natural dentro de la jerarquía de Chomsky:

- Es en gran parte libre de contexto.
- Algunos fenómenos son context-sensitive.
- Existe cierto consenso de que sólo es levemente context-sensitive.

Lenguajes No Terminalmente Separados (NTS)

Lo más cerca que conocemos del Lenguaje Natural con algún tipo de aprendibilidad (Clark, 2006):

- Subclase de lenguajes libres de contexto.
- PAC-aprendibles en su versión Inambigua (UNTS).
- Según Clark, el Lenguaje Natural no es estrictamente NTS pero está cerca.

Lenguajes No Terminalmente Separados (NTS)

Lo más cerca que conocemos del Lenguaje Natural con algún tipo de aprendibilidad (Clark, 2006):

- Subclase de lenguajes libres de contexto.
- PAC-aprendibles en su versión Inambigua (UNTS).
- Según Clark, el Lenguaje Natural no es estrictamente NTS pero está cerca.

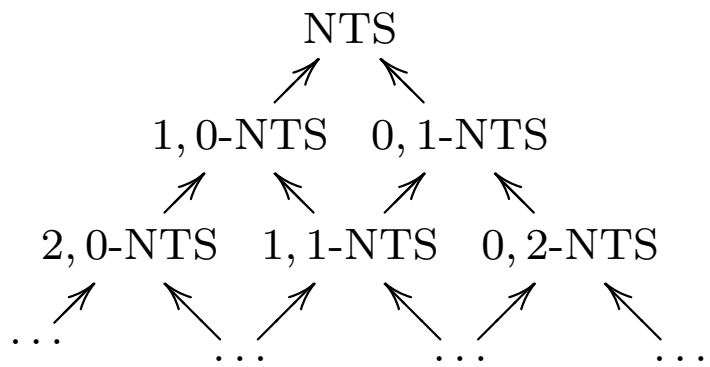
¿Cuán cerca? (Luque e Infante-Lopez, 2009 y 2010):

- Estudiamos la adecuación fuerte de las UNTS para Lenguaje Natural desde un enfoque empírico.
- Dado un treebank de evaluación, damos una forma de acotar la máxima performance F_1 obtenible con las UNTS.
- Usamos grafos con pesos y programación lineal entera (ILP).
- Las cotas dan bajas para un corpus de oraciones cortas del Inglés.

Lenguajes k, l-NTS

Buscamos una clase aprendible más adecuada para el Lenguaje Natural (Luque e Infante-Lopez, 2010):

- Generalizamos las NTS con una jerarquía de clases k, l-NTS:



- Probamos que todas las k, l-UNTS son PAC-aprendibles.
- Usamos una reducción de k, l-UNTS a UNTS con alfabeto enriquecido.
- Vemos que las cotas para $1, 1$-UNTS sobre oraciones cortas del Inglés son mucho mejores que para UNTS.

Trabajo Futuro

El algoritmo teóricamente PAC no es utilizable en la práctica. Aún para gramáticas de juguete, requieren de una gran cantidad de ejemplos para asegurar una mínima convergencia.

- Desarrollar un algoritmo práctico basado en la versión teórica.
- Incorporar heurísticas y otros elementos provenientes del área de Machine Learning.
- Realizar experimentos con gramáticas de juguete y/o gramáticas generadas aleatoriamente.
- Realizar experimentos con lenguaje natural.

¡Gracias! ¿Preguntas?