# Capítulo 8

# Análisis sintáctico estocástico

La comunicación humana depende de multitud de factores, pero todos ellos responden a una cierta regularidad y a una cierta estructura. El principal objetivo de la sintaxis dentro de la lingüística es el de intentar aislar dicha estructura. Hasta ahora, la única forma de sintaxis que permitían los métodos de etiquetación que hemos descrito es la simple consideración del orden secuencial de aparición de las palabras dentro de la frase, bien en términos de las propias palabras, o bien en términos de sus categorías léxicas. Desde este momento, nuestro propósito es escapar de la tiranía lineal impuesta por este tipo de modelos, introducir otras nociones de gramática más complejas, y comenzar a explorar su aplicación al proceso de etiquetación.

Después de introducir formalmente el concepto de gramática independiente del contexto estocástica, estudiaremos el problema del análisis sintáctico para este tipo de gramáticas. La tarea del análisis sintáctico[1] ha sido otra de las áreas de gran actividad dentro de la investigación en NLP durante los últimos años [Alonso 2000]. En el presente trabajo, no pretendemos realizar una cobertura de la viabilidad y complejidad de todas y cada una de las aproximaciones que se han desarrollado en este terreno, tal y como hicimos con las técnicas de etiquetación. En lugar de esto, comenzaremos introduciendo un sencillo e intuitivo mecanismo de análisis sintáctico, el algoritmo CYK [Kasami 1965, Younger 1967], que progresivamente iremos adaptando a nuestras necesidades.

Finalmente, veremos que al igual que los etiquetadores tienen que enfrentarse al problema de las palabras desconocidas, es decir, al problema de los diccionarios incompletos, los analizadores sintácticos deben saber enfrentarse al problema de las gramáticas incompletas. Por tanto, nuestro objetivo aquí es el de dejar preparado un marco de análisis sintáctico estocástico que permita experimentar cómodamente con técnicas de análisis sintáctico robusto[2] orientadas específicamente al problema de la etiquetación. Dichas técnicas serán descritas con detalle en el capítulo próximo.

## 8.1 Gramáticas independientes del contexto estocásticas

El modelo probabilístico más sencillo y más natural para representar las estructuras anidadas y los comportamientos recursivos de los lenguajes es quizás el de las gramáticas independientes del contexto probabilísticas, también llamadas estocásticas. Una gramática de este tipo es simplemente una gramática independiente del contexto que incorpora una probabilidad asociada a cada regla de producción. El propósito de dichas probabilidades es indicar que algunas operaciones de reescritura son más probables que otras. La única restricción impuesta por

[1] También denominada *parsing*.

[2] También denominado *robust parsing*.

este modelo es que las probabilidades de las reglas que comparten la misma parte izquierda, es decir, las reglas correspondientes a las distintas posibilidades de reescritura de un mismo símbolo, deben sumar 1. A continuación presentamos la definición formal.

**Definición 8.1** Una *gramática independiente del contexto estocástica* se define como $G = (N, T, P, S)$, donde:

- $N$ es el conjunto de *variables*, *símbolos no terminales*, o *categorías sintácticas*, es decir, símbolos que no forman parte de las frases del lenguaje generado por la gramática, pero que sirven de ayuda a la hora de describirlo.

- $T$ es el conjunto de *símbolos terminales* o *categorías léxicas*, es decir, el conjunto de los símbolos o palabras que sí forman parte de las frases del lenguaje generado por la gramática.

- $P$ es el conjunto de *producciones* o *reglas de reescritura* de la forma $A \to \alpha$, donde $A \in N$, es decir, es un símbolo no terminal, y $\alpha \in (N \cup T)^*$, es decir, es cualquier combinación de cero, uno o más símbolos terminales y no terminales. Cada regla tiene asociada una probabilidad, y este conjunto de probabilidades verifica:

$$\sum_{\alpha} P(A \to \alpha) = 1, \qquad \forall A \in N. \tag{8.1}$$

- $S$ es un elemento destacado del conjunto $N$, que se denomina *axioma* o *símbolo inicial*. Todas las frases pertenecientes al lenguaje generado por la gramática han de tener un árbol de análisis cuya raíz debe ser el símbolo $S$.

Es importante señalar que cuando escribimos $P(A \to \alpha)$, lo que realmente queremos decir es $P(A \to \alpha|A)$. Por tanto, para cada símbolo no terminal $A$, la gramática proporciona la distribución de probabilidad de todas sus posibles transformaciones $\alpha$. □

El método para diseñar una gramática estocástica, es decir, la manera de identificar los símbolos, las producciones y las probabilidades, puede ser manual cuando la gramática es pequeña. Pero en la práctica, para las gramáticas de los lenguajes naturales, tal y como se discutió ya en la sección 2.3.2, todos estos elementos se suelen extraer automáticamente de recursos lingüísticos especializados en forma de bancos de árboles[3].

Una gramática estocástica es un formalismo que describe o genera un lenguaje. El *lenguaje generado por una gramática G* se denota por $L(G)$. Asociado a dicho formalismo generador existen, como veremos más adelante, algoritmos para verificar si una determinada frase $s$ pertenece o no a $L(G)$. Estos algoritmos son la base de los analizadores sintácticos[4], los cuales deben ser capaces también de obtener el árbol de análisis para dicha frase $s$, cuando $s \in L(G)$. Podría ocurrir incluso que dicho árbol de análisis no fuera único, en cuyo caso se dice que la gramática es *ambigua*. Al analizar una frase, el analizador debe ser capaz de obtener todos sus posibles árboles de análisis.

Por el momento, nos interesa centrarnos en el problema de cómo asignar una probabilidad a cada frase. La probabilidad de una frase $s$, de acuerdo con una gramática $G$, viene dada por:

$$P(s) = \sum_{t} P(s,t) = \sum_{t\,:\,\mathrm{hojas}(t)=s} P(t),$$

[3] O *treebanks*.
[4] O también *parsers*.

donde la variable $t$ recorre el espacio de todos los posibles árboles de análisis para los cuales la secuencia de nodos hoja, leída de izquierda a derecha, coincide con la frase $s$. Asumiendo la hipótesis de que las reglas de una gramática estocástica $G$ son independientes, la probabilidad de un nodo cualquiera de un árbol $t$ se calcula recursivamente como el producto de las probabilidades de sus subárboles locales y de la probabilidad de la regla de producción de $G$ que los une. La probabilidad de un árbol $t$ viene dada, por tanto, por la probabilidad de su nodo raíz.

**Ejemplo 8.1** Sea $G = (N, T, P, S)$ una gramática independiente del contexto estocástica, donde $N = \{S, A, B, C\}$, $T = \{a, b\}$, el conjunto de reglas $P$, con sus respectivas probabilidades entre parántesis, viene dado por:

$$\begin{array}{llllllll} S \to A\,B & (0,25), & A \to B\,A & (0,5), & B \to C\,C & (0,1), & C \to A\,B & (0,2), \\ S \to B\,C & (0,75), & A \to a & (0,5), & B \to b & (0,9), & C \to a & (0,8), \end{array}$$

y $S = S$. La frase $s = b\,b\,a\,b$ tiene dos posibles árboles de análisis, tal y como se muestra en la figura 8.1. En esta figura, los símbolos no terminales de cada nodo llevan como subíndice la probabilidad de la regla mediante la cual generan los subárboles que encabezan. Así pues, la probabilidad de cada árbol es:

$$P(t_1) = 0,9 \times 0,9 \times 0,5 \times 0,5 \times 0,5 \times 0,9 \times 0,25 = 0,02278.$$
$$P(t_2) = 0,9 \times 0,9 \times 0,5 \times 0,5 \times 0,9 \times 0,2 \times 0,75 = 0,02733.$$

Y por tanto, $P(s) = P(t_1) + P(t_2) = 0,02733 + 0,02278 = 0,05011$. □

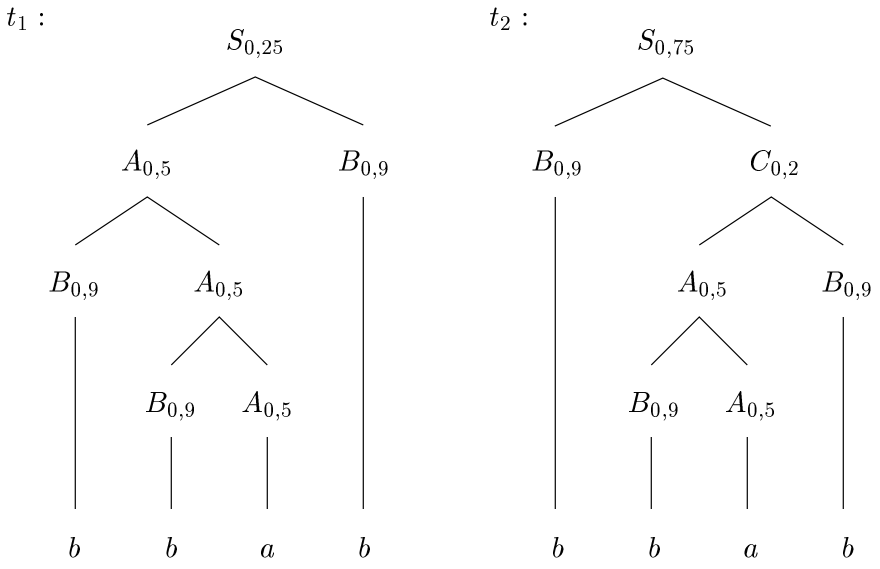


Figura 8.1: Árboles de análisis para la frase $s = b\,b\,a\,b$

### 8.1.1 Algunas características de las gramáticas estocásticas

A continuación citamos algunas circunstancias en las que puede ser conveniente el uso de gramáticas estocásticas, y también algunas ideas sobre sus limitaciones:

- A medida que las gramáticas se expanden para conseguir la mayor cobertura posible sobre grandes colecciones de textos, la ambigüedad crece también con ellas. Como ya sabemos, este fenómeno da como resultado la existencia de múltiples análisis estructurales diferentes para una misma secuencia de palabras. Con el uso de gramáticas estocásticas, se puede obtener una cierta idea de la plausibilidad de cada uno de esos análisis. La gramática del ejemplo 8.1 no sólo permite calcular la probabilidad de una frase, en este caso $s = b\,b\,a\,b$, sino que además nos indica cuál de los análisis es el más probable, en este caso $t_2$.

- Como hemos visto, el procedimiento más seguro para la construcción de una gramática para un lenguaje natural es la extracción de reglas a partir de un banco de árboles. Sin embargo, existen métodos capaces de realizar inferencia gramatical[5] sobre textos sin ningún tipo de marcación sintáctica. Aunque este tipo de inferencia gramatical *desde cero* es una difícil tarea todavía sin resolver, parece que el aprendizaje de gramáticas independientes del contexto no se puede realizar sin evidencias negativas, es decir, sin una provisión de ejemplos gramaticalmente incorrectos [Gold 1967], mientras que las gramáticas independientes del contexto estocásticas sí se pueden generar con sólo datos positivos [Horning 1969].

- Las gramáticas estocásticas presentan un buen compromiso de robustez. Los textos reales tienden a reflejar los errores léxicos y sintácticos más comunes de los hablantes. La manera obvia de evitar este problema es identificar las frases en las cuales se localizan los errores y eliminarlas del proceso de extracción de reglas. Sin embargo, se puede prescindir de esa tarea, y dejar que una gramática estocástica asigne de manera natural una probabilidad baja a las frases menos plausibles.

- En la práctica, existen idiomas para los cuales los $n$-gramas de palabras ($n > 1$) podrían constituir un modelo de lenguaje mejor que el representado por las gramáticas estocásticas. Un modelo de $n$-gramas de palabras tiene en cuenta dependencias contextuales entre elementos concretos del léxico que, en general, las gramáticas no utilizan.

- Las gramáticas estocásticas no parecen ser del todo *imparciales* en algunos aspectos, lo cual puede resultar inadecuado para determinadas aplicaciones. Por ejemplo, en general, la probabilidad de un árbol pequeño es mayor que la de un árbol grande. Esto podría no ser demasiado importante, ya que las frases de un lenguaje tienden a tener una cierta longitud intermedia. Pero no cabe duda de que una gramática estocástica asigna demasiada masa de probabilidad a las frases más cortas. De igual manera, en los árboles de análisis, los símbolos no terminales con un número bajo de posibles reescrituras se ven favorecidos sobre los no terminales con muchas posibilidades, ya que las reglas individuales de estos últimos tendrán probabilidades mucho más bajas.

- Por último, es importante comentar que no está claro que la sintaxis de todos los lenguajes naturales encaje dentro del marco de las gramáticas independientes del contexto, estocásticas o no estocásticas. Incluso aunque lo hiciera, el formalismo se queda muy justo y presenta limitaciones. No obstante, a pesar de su simplicidad, las gramáticas independientes del contexto todavía permiten expresar una gran variedad de las construcciones sintácticas que aparecen en la mayoría de los idiomas, y llevan asociados algoritmos muy eficientes para múltiples tareas de comprensión del lenguaje, una de las cuales es el análisis sintáctico, como veremos más adelante.

En definitiva, lo importante es que las gramáticas estocásticas proporcionan modelos probabilísticos del lenguaje. En un primer momento, cabría esperar, por tanto, que si todas las reglas de producción verifican la restricción (8.1), entonces

$$\sum_{s \in L(G)} P(s) = \sum_t P(t) = 1.$$

[5] *Grammar induction.*

Realmente, esto es cierto sólo si la masa de probabilidad de las reglas se acumula en un número finito de árboles de análisis. Así pues, consideremos el siguiente ejemplo.

**Ejemplo 8.2** Sea $G$ una gramática estocástica con un único símbolo no terminal $S$, un único símbolo terminal $a$, y un conjunto de reglas:

$$\begin{array}{ll} S \rightarrow a & (\frac{1}{3}), \\ S \rightarrow S\ S & (\frac{2}{3}). \end{array}$$

Esta gramática genera frases de la forma $a$, $aa$, $aaa$, ... Sin embargo, las probabilidades de estas frases son de la forma:

$$\begin{array}{lcl} P(a) & = & \frac{1}{3} \\ P(aa) & = & \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{2}{27} \\ P(aaa) & = & (\frac{2}{3})^2 \times (\frac{1}{3})^3 \times 2 = \frac{8}{243} \\ & \vdots & \end{array}$$

La probabilidad del lenguaje $L(G)$ es la suma de la serie infinita $\frac{1}{3} + \frac{2}{27} + \frac{8}{243} + \ldots$, la cual tiende a $\frac{1}{2}$. Por tanto, la mitad de la masa de probabilidad ha desaparecido en el conjunto infinito de árboles que no generan frases de este lenguaje. □

Distribuciones de probabilidad como la del ejemplo anterior se denominan normalmente *distribuciones inconsistentes*. En la práctica, el uso de distribuciones inconsistentes no presenta excesivos problemas. A menudo, ni siquiera importa si la distribución es consistente o no, especialmente cuando nuestro objetivo principal es la comparación de las magnitudes de probabilidad de los diferentes análisis. Además, Chi y Geman demuestran que si los parámetros de nuestras gramáticas estocásticas se estiman a partir de bancos de árboles, siempre podemos obtener distribuciones de probabilidad consistentes [Chi y Geman 1998].

### 8.1.2 Relación entre gramáticas estocásticas y HMM,s

Las mismas tres preguntas fundamentales que planteamos en la sección 4.4 para los HMM,s son también aplicables a las gramáticas estocásticas. De hecho, un HMM se puede ver como una gramática regular estocástica. En el caso de las gramáticas estocásticas, las preguntas son las siguientes:

1. Dada una frase $s$ y dada una gramática $G$, ¿cuál es la probabilidad $P(s|G)$, es decir, la probabilidad de la frase $s$ de acuerdo con la gramática $G$?

2. ¿Cuál es el $\arg\max_t P(t|s, G)$, es decir, el árbol de análisis más probable para la frase $s$?

3. ¿Cómo podemos elegir el $\arg\max_G P(s|G)$, es decir, las probabilidades de las reglas de $G$ que maximizan la probabilidad de una determinada frase $s$?

En relación con la primera pregunta, hemos visto que la probabilidad de una frase de acuerdo con una gramática se puede calcular como la suma de las probabilidades de todos sus árboles de análisis. Sin embargo, desafortunadamente, el número de análisis de una frase puede crecer exponencialmente con la longitud de la frase, de forma que la simple suma de las probabilidades de esos análisis no constituye un buen método. Existen algoritmos especializados para realizar de manera más eficiente este cálculo, tales como el *algoritmo de las probabilidades externas* y el *algoritmo de las probabilidades internas*. Estos algoritmos se apoyan, respectivamente, en la definición de las probabilidades externas de un nodo de un árbol como

$$\alpha_j(p,q) = P(w_{1(p-1)}, N^j_{pq}, w_{(q+1)n}|G)$$

y de las probabilidades internas como

$$\beta_j(p,q) = P(w_{pq}|N^j_{pq}, G)$$

donde $w_{ik}$ es la secuencia de palabras $w_i, w_{i+1}, \ldots, w_k$ y $N^j_{pq}$ es un subárbol encabezado por el símbolo no terminal $N^j$ que produce la secuencia de palabras $w_{pq}$. El algoritmo de las probabilidades externas es un algoritmo de programación dinámica similar al *algoritmo hacia adelante* de los HMM,s, y el algoritmo de las probabilidades internas es similar al *algoritmo hacia atrás*.

Para resolver el problema que se plantea en la segunda pregunta, existe también un algoritmo especializado. Una vez más, la estrategia para encontrar el análisis más probable de una frase es similar al algoritmo que busca el camino más probable a través de un HMM que produce dicha frase (el algoritmo de Viterbi). El método para gramáticas estocásticas es una variante del algoritmo de las probabilidades internas, que busca el elemento que maximiza la suma y recuerda qué regla dio lugar a ese máximo. Al igual que en el caso de un HMM, este método funciona gracias a la hipótesis de independencia de las reglas gramaticales. El resultado es un algoritmo de análisis sintáctico de complejidad $\mathcal{O}(n^3m^3)$, donde $n$ es el número de palabras de la frase a analizar y $m$ es el número de símbolos no terminales de la gramática.

Por último, tal y como sugiere la tercera pregunta, existen algoritmos para el *entrenamiento* de gramáticas estocásticas. La idea de entrenamiento aquí es la misma que la del aprendizaje o inferencia gramatical, pero en un sentido limitado. Se asume que la estructura de la gramática, es decir, los conjuntos de símbolos terminales, no terminales y reglas, se conoce de antemano. El entrenamiento de la gramática es entonces simplemente un proceso que intenta optimizar las probabilidades de las reglas. Para determinar estas probabilidades respetando la restricción (8.1), nos gustaría calcular $P(A \to \alpha)$ como

$$P(A \to \alpha) = \frac{\text{número de veces que aparece la regla } A \to \alpha}{\displaystyle\sum_{\beta} \text{número de veces que aparece la regla } A \to \beta}$$

donde $A$ es cualquier símbolo no terminal y $\alpha$ y $\beta$ son cualquier combinación de símbolos terminales y no terminales. Si tenemos disponible un corpus anotado sintácticamente, es decir, un banco de árboles, las probabilidades se pueden calcular directamente (como se discutió en la sección 2.3.2). Si no tenemos un corpus anotado disponible, al igual que en el caso de los HMM,s, se puede construir un algoritmo de entrenamiento EM[6], que toma una gramática estocástica inicial y un corpus no anotado, y ajusta las probabilidades de las reglas de la gramática de manera que las frases de dicho corpus obtengan la máxima probabilidad posible. Las limitaciones de los métodos de entrenamiento para HMM,s también están presentes aquí, pues sólo garantizan que se encuentren máximos locales [Charniak 1993].

La descripción de todas estas estrategias desborda el ámbito del presente trabajo. Una excelente introducción a los mismos puede verse en [Manning y Schütze 1999, cap. 11]. A partir de ahora preferimos centrar nuestro interés en los algoritmos de análisis sintáctico, los cuales, una vez adaptados al marco estocástico, también pueden proporcionar en sí mismos soluciones a los problemas planteados en las preguntas 1 y 2 de una manera más intuitiva.

[6] *Expectation-Maximization* (maximización de la esperanza), cuya parte E combina las probabilidades internas y externas dando lugar al algoritmo *inside-outside* (hacia adentro y hacia afuera).