Apéndice B

Algoritmos de análisis sintáctico para CFG

En este apéndice se describen los esquemas de análisis sintáctico de aquellos algoritmos de análisis de gramáticas independientes del contexto que han servido como base para la obtención de algoritmos de análisis sintáctico para LIG y TAG. En consecuencia, se abordan los algoritmos CYK, Earley ascendente y Earley.

B.1 El algoritmo CYK

El algoritmo de Cocke-Younger-Kasami (CYK) para análisis sintáctico de gramáticas independientes del contexto en forma normal de Chomsky [85] fue descubierto por J. Cocke, pero fue publicado independientemente por Younger [234] y Kasami [97], de ahí su nombre.

Es este un algoritmo ascendente puro basado en programación dinámica. A este respecto, la versión original del algoritmo hace uso de una matriz bidimensional indexada por posiciones de la cadena de entrada para almacenar los resultados parciales obtenidos, de tal modo que el elemento A se encuentra en la posición $[i, j]$ de dicha matriz si y sólo si $A \stackrel{*}{\Rightarrow} a_{i+1} \dots a_j$. Sin embargo, seguiremos aquí el enfoque de Sikkel [189] de abstraer las estructuras de datos utilizadas puesto que estas forman parte de los detalles de implementación y no son esenciales al algoritmo. En este contexto, consideraremos que el algoritmo CYK construye un conjunto de ítems

$$\left\{ \ [A, i, j] \mid A \stackrel{*}{\Rightarrow} a_{i+1} \dots a_n \ \right\}$$

Una cadena de entrada $a_1 \dots a_n$ pertenece al lenguaje generado por la gramática si y sólo si el ítem $[S, 0, n]$ se encuentra en dicho conjunto.

Esquema de análisis sintáctico B.1 El sistema de análisis sintáctico $\mathbb{P}_{\mathrm{CYK}}$ correspondiente al algoritmo CYK para una CFG $\mathcal{G} = (V_N, V_T, P, S)$ en forma normal de Chomsky, esto es, con producciones de la forma $A \to BC$ o $A \to a$, donde $A, B, C \in V_N$ y $a \in V_T$, y una cadena de entrada $a_1 \dots a_n$ es el que se muestra a continuación:

$$\mathcal{I}_{\mathrm{CYK}} = \left\{ \ [A, i, j] \mid A \in V_N, \ 0 \leq i \leq j\} \ \right\}$$

$$\mathcal{H}_{\mathrm{CYK}} = \left\{ \ [a, i-1, i] \mid a = a_i \ \right\}$$

$$\mathcal{D}_{\mathrm{CYK}}^{\mathrm{Scan}} = \left\{ \ [a, i, i+1] \vdash [A, i, i+1] \mid A \to a \in P \ \right\}$$

$$\mathcal{D}_{\mathrm{CYK}}^{\mathrm{Comp}} = \{\ [B,i,k],[C,k,j] \vdash [A,i,j] \mid A \rightarrow BC \in P\ \}$$

$$\mathcal{D}_{\mathrm{CYK}} = \mathcal{D}_{\mathrm{CYK}}^{\mathrm{Scan}} \cup \mathcal{D}_{\mathrm{CYK}}^{\mathrm{Comp}}$$

$$\mathcal{F}_{\mathrm{CYK}} = \{\ [S,0,n]\ \}$$

§

Básicamente, podemos decir que el algoritmo CYK comienza por la creación de todos los ítems que se corresponden con reglas terminales (pasos $\mathcal{D}_{\mathrm{CYK}}^{\mathrm{Scan}}$) para posteriormente tratar de combinar en parejas los ítems generados a fin de reconocer las reglas binarias (pasos $\mathcal{D}_{\mathrm{CYK}}^{\mathrm{Comp}}$). El algoritmo termina cuando no se pueden combinar más ítems. En tal caso, si se ha generado algún ítem final, la cadena de entrada ha sido reconocida.

B.2 Una versión ascendente del algoritmo de Earley

El algoritmo CYK presenta una importante limitación, ya que sólo es aplicable a gramáticas en forma normal de Chomsky. Nuestro objetivo ahora es extender el esquema CYK a la clase general de CFG, obteniendo un esquema Earley ascendente [191] en el que se construye un conjunto de ítems

$$\left\{\ [A \rightarrow \alpha \bullet \beta, i, j] \mid \alpha \stackrel{*}{\Rightarrow} a_{i+1} \ldots j\ \right\}$$

que nos permitirán representar el reconocimiento parcial de producciones mediante la utilización de producciones con punto, al contrario de lo que ocurría en el caso del algoritmo CYK, el cual sólo permitía representar producciones binarias completas. En este sentido, un ítem CYK $[A,i,j]$ puede ser equivalentemente representado por $[A \rightarrow \alpha\bullet, i, j]$, donde $\alpha = BC$ o bien $\alpha = a_i$. El punto en las producciones indica que los elementos gramaticales situado a su izquierda han sido reconocidos. Las producciones son por tanto reconocidas de izquierda a derecha de tal modo que cuando el punto está a la derecha del último elemento del lado derecho de una producción si y sólo si esta ha sido completamente reconocida.

Esquema de análisis sintáctico B.2 El sistema de análisis $\mathbb{P}_{\mathrm{buE}}$ correspondiente al algoritmo Earley ascendente para una gramática independendiente del contexto $\mathcal{G}$ y una cadena de entrada $a_1 \ldots a_n$ es el siguiente:

$$\mathcal{I}_{\mathrm{buE}} = \{\ [A \rightarrow \alpha \bullet \beta, i, j] \mid A \rightarrow \alpha\beta \in P,\ 0 \leq i \leq j\ \}$$

$$\mathcal{H}_{\mathrm{buE}} = \mathcal{H}_{\mathrm{CYK}}$$

$$\mathcal{D}_{\mathrm{buE}}^{\mathrm{Init}} = \{\ \vdash [A \rightarrow \bullet\alpha, i, i]\ \}$$

$$\mathcal{D}_{\mathrm{buE}}^{\mathrm{Scan}} = \{\ [A \rightarrow \alpha \bullet a\beta, i, j], [a, j, j+1] \vdash [A \rightarrow \alpha a \bullet \beta, i, j+1]\ \}$$

$$\mathcal{D}_{\text{buE}}^{\text{Comp}} = \left\{ \ [A \to \alpha \bullet B\beta, i, k], [B \to \gamma \bullet, k, j] \vdash [A \to \alpha B \bullet \beta, i, j] \ \right\}$$

$$\mathcal{D}_{\text{buE}} = \mathcal{D}_{\text{buE}}^{\text{Init}} \cup \mathcal{D}_{\text{buE}}^{\text{Scan}} \cup \mathcal{D}_{\text{buE}}^{\text{Comp}}$$

$$\mathcal{F}_{\text{buE}} = \left\{ \ [S \to \alpha \bullet, 0, n] \ \right\}$$

§

Proposición B.1 CYK $\overset{\text{ir}}{\Longrightarrow}\overset{\text{sr}}{\Longrightarrow}\overset{\text{ext}}{\Longrightarrow}$ buE .

Efectivamente, el sistema de análisis $\mathbb{P}_{\text{buE}}$ se deriva del sistema de análisis $\mathbb{P}_{\text{CYK}}$ mediante la aplicación de un refinamiento de los ítems y de un refinamiento de pasos deductivos, puesto que se ha descompuesto el paso $\mathcal{D}_{\text{CYK}}^{\text{Comp}}$ en dos pasos $\mathcal{D}_{\text{buE}}^{\text{Init}}$ y $\mathcal{D}_{\text{buE}}^{\text{Comp}}$. Finalmente se ha realizado una extensión del sistema de análisis al considerar, no ya gramáticas en forma normal de Chomsky, sino en forma arbitraria. La prueba formal puede encontrarse en [189].

Con respecto al comportamiento del algoritmo, podemos resumirlo indicando que el análisis comienza con la creación por parte de los pasos $\mathcal{D}_{\text{bu}}^{\text{Init}}$ de los ítems $[A \to \bullet\alpha, i, i]$ para toda regla de la gramática y para toda posición en la cadena de entrada. A continuación se aplican los pasos $\mathcal{D}_{\text{buE}}^{\text{Scan}}$ y $\mathcal{D}_{\text{buE}}^{\text{Comp}}$ con el fin de ir desplazando el punto de las producciones hacia la derecha.

Un paso deductivo $\mathcal{D}_{\text{buE}}^{\text{Scan}}$ es aplicable a ítems de la forma $[A \to \alpha \bullet a\beta, i, j]$ cuando $a_{j+1} = a$, obteniéndose un nuevo ítem $[A \to \alpha a \bullet \beta, i, j+1]$. Es decir, se ha reconocido el símbolo terminal que estaba justo a la derecha del punto.

Un paso deductivo $\mathcal{D}_{\text{buE}}^{\text{Comp}}$ se aplica cuando un ítem tiene una regla con el punto en el extremo derecho. Dado un ítem $[B \to \gamma\bullet, k, j]$ se buscan todos los posibles $[A \to \alpha \bullet B\beta, i, k]$ y se generan nuevos ítems $[A \to \alpha B \bullet \beta, i, j]$ que representan que la subcadena $a_{k+1} \dots a_j$ puede ser reducida a B y por consiguiente, como la subcadena $a_{i+1} \dots a_k$ se reduce a α, la subcadena $a_{i+1} \dots a_j$ se reduce a αB.

El proceso termina cuando no se pueden combinar más ítems. En tal caso, si se ha generado algún ítem final la cadena de entrada pertenece al lenguaje generado por la gramática.

B.3 El algoritmo de Earley

Se puede derivar un esquema de análisis correspondiente al algoritmo de Earley [69] a partir del algoritmo de Earley ascendente mediante la aplicación de un *filtrado dinámico* a este último, de modo que los ítems de la forma $[A \to \bullet\alpha, i, i]$ no sean generados por los pasos Init para todas las posibles producciones y posiciones en la cadena de entrada, sino que sean generados o no dependiendo de la validez de otros items mediante un paso deductivo predictivo, encargándose el paso Init únicamente de la creación de los ítems correspondientes a las producciones del axioma de la gramática. Los ítems válidos son por tanto de la forma

$$\left\{ \ [A \to \alpha \bullet \beta, i, j] \mid \alpha \overset{*}{\Rightarrow} a_{i+1} \dots j, \ S \overset{*}{\Rightarrow} a_1 \dots a_i A\delta \ \right\}$$

Esquema de análisis sintáctico B.3 El sistema de análisis $\mathbb{P}_{\text{Earley}}$ correspondiente al algoritmo de Earley para una gramática independiente del contexto $\mathcal{G}$ y una cadena de entrada $A_1 \dots a_n$ es el que se muestra a continuación:

$$\mathcal{I}_{\text{Earley}} = \mathcal{I}_{\text{buE}}$$

$$\mathcal{H}_{\text{Earley}} = \mathcal{H}_{\text{CYK}}$$

$$\mathcal{D}^{\text{Init}}_{\text{Earley}} = \left\{ \ \vdash [S \to \bullet\alpha, 0, 0] \ \right\}$$

$$\mathcal{D}^{\text{Scan}}_{\text{Earley}} = \mathcal{D}^{\text{Scan}}_{\text{buE}}$$

$$\mathcal{D}^{\text{Pred}}_{\text{Earley}} = \left\{ \ [A \to \alpha \bullet B\beta, i, j] \vdash [B \to \bullet\gamma, j, j] \ \right\}$$

$$\mathcal{D}^{\text{Comp}}_{\text{Earley}} = \mathcal{D}^{\text{Comp}}_{\text{buE}}$$

$$\mathcal{D}_{\text{Earley}} = \mathcal{D}^{\text{Init}}_{\text{Earley}} \cup \mathcal{D}^{\text{Scan}}_{\text{Earley}} \cup \mathcal{D}^{\text{Pred}}_{\text{Earley}} \cup \mathcal{D}^{\text{Comp}}_{\text{Earley}}$$

$$\mathcal{F}_{\text{Earley}} = \mathcal{F}_{\text{buE}}$$

§

Proposición B.2 buE $\stackrel{\text{df}}{\Longrightarrow}$ Earley .

La prueba de dicha transformación puede encontrase en [189]. Con respecto al comportamiento del algoritmo, lo resumiremos indicando que el proceso de análisis comienza con la generación, por parte del conjunto de pasos $\mathcal{D}^{\text{Init}}_{\text{Earley}}$, de los ítems de la forma $[S \to \bullet\alpha, 0, 0]$, donde $S \to \alpha \in P$ es una regla del axioma de la gramática. Estos ítems indican que se está tratando reconocer el axioma de la gramática desde el inicio de la cadena de entrada. Los pasos deductivos $\mathcal{D}^{\text{Scan}}_{\text{Earley}}$ y $\mathcal{D}^{\text{Comp}}_{\text{Earley}}$ se comportan como en el caso del algoritmo de Earley ascendente, mientras que el conjunto de pasos $\mathcal{D}^{\text{Pred}}_{\text{Earley}}$ se encarga de realizar la fase descendente o predictiva del algoritmo, ya que a partir de ítems de la forma $[A \to \alpha \bullet B\beta, i, j]$ genera los ítems $[B \to \bullet\gamma, j, j]$, esto es, se predicen todas las reglas que potencialmente pueden ser útiles en el reconocimiento de la cadena de entrada.