El Algoritmo SCP de Análisis Sintáctico mediante Propagación de Restricciones

José F. Quesada
CICA (Centro de Informática Científica de Andalucía)
josefran@cica.es

Directores: Dr. D. José Gabriel de Amores Carredano y Dr. D. José Antonio Alonso Jiménez. **Departamento**: de Algebra, Computación, Geometría y Topología, Universidad de Sevilla. **Tribunal**: Dr. D. Rafael Infante Macías, Dr. D. Mario de Jesús Pérez Jiménez, Dra. D^{a}. María Felisa Verdejo Maillo, Dr. D. José Meseguer Guaita, y Dr. D. Horacio Rodríguez Hontoria. **Fecha de Lectura**: 17 de Junio de 1997. **Calificación**: *Apto cum laude* por unanimidad. **Título de Doctor obtenido**: Doctor en Informática.

El análisis sintáctico (parsing) de gramáticas libres de contexto (GLC) ha sido y continúa siendo uno de los campos de investigación más activos tanto en PLN como en Informática (Compiladores). En este ámbito se pueden situar los algoritmos de Knuth y DeRemer para el análisis de lenguajes LR, SLR y LALR. En los últimos años el desarrollo de compiladores se ha interesado por fenómenos expresivamente más potentes (modelos de especificación algebraica, sistemas basados en lógica de reescritura (Meseguer), etc.), lo que exige obtener algoritmos muy robustos y eficientes para GLC.

Por otro lado, en el campo del PLN el interés por los fenómenos LC ha propiciado la aparición de múltiples algoritmos entre los que destacamos el CKY de Cocke, Kasamy y Younger, el algoritmo de Earley, las técnicas tabulares o basadas en charts originadas a partir de la propuesta de Kay, los diferentes algoritmos basados en programación lógica sobre los que descansan otras muchas gramáticas lógicas como DCG, el famoso algoritmo GLR de Tomita, la utilización de técnicas probabilísticas, paralelismo masivo o incluso técnicas reduccionistas como la aproximación de GLC mediante autómatas de estados finitos o la poda y especialización gramatical. Asimismo la mayor parte de los formalismos y teorías gramaticales basados en unificación utilizan un núcleo LC ampliado o extendido con ecuaciones funcionales.

La tesis se incardina en esta tradición, y propone un nuevo algoritmo (SCP) para el análisis de GLC. La primera parte de la tesis presenta la motivación para el algoritmo desde tres puntos de vista: formal, computacional y lingüístico. Desde el enfoque formal, el algoritmo

SCP intenta incorporar *inteligencia* al proceso de análisis sintáctico, eliminando lo que denominamos ambigüedad o sobregeneración del analizador (por oposición a la ambigüedad gramatical). Este modelo de *inteligencia formal* es captado mediante las relaciones de derivabilidad parcial, adyacencia y cobertura. Desde el punto de vista computacional se proponen los multi árboles virtuales para la representación de los árboles o bosques de análisis. Finalmente, desde el punto de vista lingüístico, el algoritmo está diseñado para adaptarse a los fenómenos más habituales de los lenguajes naturales.

La segunda parte de la tesis presenta el algoritmo en sí, lo que incluye la presentación del modelo $Q-mem$ para la compilación y representación de GLC, y el modelo $RB-mem$ basado en bloques reusables de memoria para la gestión de la memoria.

Finalmente, la tercera parte analiza las consecuencias y propiedades tanto lingüísticas, como computacionales y formales del algoritmo. A nivel lingüístico se lleva a cabo un recorrido sistemático por todos los fenómenos descritos en la literatura especializada. La consecuencia básica es que el algoritmo SCP permite analizar correctamente todos los fenómenos. Comparado con el resto de algoritmos (CKY, Earley, chart, GLR, etc.) el algoritmo SCP siempre muestra un nivel inferior de sobregeneración, que llega a ser nulo en 15 de los 17 fenómenos descritos. Entre los fenómenos analizados se encuentran la recursividad, dependencias locales y no locales, gramáticas cíclicas, ε-cíclicas, L-epsilon (*left-hyde recursion*), RL-epsilon, ambigüedad exponencial, tratamiento de la sentencia vacía, producciones nulas, sentencias laberínticas, etc.

Desde el punto de vista computacional, el modelo formal y los módulos $Q-mem$ y $RB-mem$ garantizan niveles de eficiencia real que multiplican entre 100 y 1000 veces los resultados descritos en la literatura. En concreto, la implementación del algoritmo ha mostrado un comportamiento estable con un rendimiento de entre 5000 y 20000 palabras analizadas por segundo, dependiendo del tipo de fenómeno. Asimismo el estudio de la complejidad computacional ha obtenido un modelo preciso de las nociones de ambigüedad, profundidad, densidad y conexión gramaticales.

Finalmente, la tesis incluye la demostración de los teoremas de corrección y completitud del algoritmo SCP. Esta demostración se construye a partir de las nociones de partición, π-equivalencia, cubrimiento, cubrimiento localmente conexo, LC-derivación y forma conexa.

Resumiendo, desde un punto de vista formal sólo el algoritmo de Earley es comparable al algoritmo SCP. El resto de algoritmos necesitan serias (y muy ineficientes) modificaciones para lograr la completitud. Por otro lado, el algoritmo SCP ha obtenido mejores resultados (eficiencia real obtenida mediante experimentos) que el resto de algoritmos, multiplicando en 2 a 3 órdenes de magnitud los resultados descritos en la literatura. Finalmente a nivel lingüístico, el algoritmo SCP muestra una especial adaptación a los fenómenos propios de los lenguajes naturales.