Procesamiento del Lenguaje Natural, num. 32 (2004), pp.75-76 recibido 6-12-2003; aceptado 20-2-2004

Análisis Sintáctico Eficiente de Gramáticas de Adjunción e Inserción de Árboles

Vicente Carrillo
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Avda. Reina Mercedes s/n
41012 Sevilla
carrillo@lsi.us.es

Resumen: Tesis doctoral en Informática realizada por Vicente Carrillo Montero bajo la dirección del doctor Víctor J. Díaz Mdrigal (Univ. de Sevilla). El acto de defensa de la tesis tuvo lugar el 17 de julio de 2003 ante el tribunal formado por los doctores Miguel Toro Bonilla (Univ. de Sevilla), Manuel Palomar Sanz (Univ. de Alicante), Horacio Rodríguez Hontoria (Univ. Politécnica de Cataluña), Alfonso Ureña López (Univ. de Jaén) y Miguel A. Alonso Pardo (Univ. de La Coruña). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: análisis sintáctico, gramáticas de adjunción de árboles, gramáticas de inserción de árboles

Abstract: PhD Thesis in Computer Science written by Vicente Carrillo Montero under the supervision of Dr. Víctor J. Díaz Madrigal (Univ. de Sevilla). The author was examined in July 17^{th}, 2003 by the commitee formed by Dr. Miguel Toro Bonilla (Univ. de Sevilla), Dr. Manuel Palomar Sanz (Univ. de Alicante), Dr. Horacio Rodríguez Hontoria (Univ. Politécnica de Cataluña), Dr. Alfonso Ureña López (Univ. de Jaén) and Dr. Miguel A. Alonso Pardo (Univ. de La Coruña). The grade obtained was *Sobresaliente Cum Laude*.
Keywords: parsing, tree adjoining grammars, tree insertion grammars

1 Introducción

Las gramáticas de adjunción de árboles (TAG, *Tree Adjoining Grammars*) es un formalismo gramatical que presenta características muy adecuadas para la descripción de lenguajes naturales. La mayoría de los analizadores tabulares definidos para ellas son adaptaciones de analizadores tabulares conocidos para gramáticas independientes del contexto (CFG, *Context Free Grammars*), sin embargo, el coste computacional teórico en el caso peor de estos algoritmos es bastante más elevado. Así, en el caso peor, los algoritmos de análisis sintáctico para TAG presentan una complejidad de $\mathcal{O}(n^6)$ en tiempo y $\mathcal{O}(n^4)$ en espacio, donde n es la longitud de la cadena a analizar, frente a $\mathcal{O}(n^3)$ en tiempo y $\mathcal{O}(n^2)$ en espacio en los analizadores para CFG.

Las gramáticas de inserción de árboles (TIG, *Tree Insertion Grammars*) es un compromiso entre CFG y TAG. Las TIG se pueden analizar con el mismo coste que las CFG, manteniendo las propiedades lingüísticas fundamentales de las TAG. Como contrapartida, su potencia expresiva queda reducida a lenguajes independientes del contexto. La importancia de este formalismo radica en que la mayor parte de las estructuras que definen los lenguajes naturales se adaptan a él. Al igual que en las TAG, los analizadores tabulares definidos para TIG están basados en los de las CFG, aunque el número de ellos es significativamente menor que el de los analizadores para TAG, debido fundamentalmente a su corta historia.

Los esquemas de análisis sintáctico establecen un método general para la descripción de algoritmos de análisis sintáctico y constituyen la formalización de estudios previos sobre analizadores deductivos. Permiten definir, analizar y relacionar algoritmos de análisis, y se han empleado satisfactoriamente con un gran número de analizadores para CFG. Recientemente se han desarrollado trabajos que usan los esquemas de análisis para describir y relacionar analizadores tabulares para TAG.

ISSN 1135-5948

2 *Análisis eficiente de TAG*

Partimos de un conjunto de analizadores tabulares para TAG definidos mediante esquemas de análisis sintáctico y derivados a partir de algoritmos conocidos para CFG con diversas estrategias. Concretamente, los esquemas basados en los algoritmos con estrategia ascendente **CYK**, **buE** y **dVH**, y los de estrategia ascendente predictiva **E** y **Nederhof**. Estos esquemas de análisis se organizan como una red de analizadores en base a las relaciones formales existentes entre ellos, la cual establece cómo se pueden derivar unos esquemas a partir de otros. En cuanto al formalismo TIG, sólo se disponía de un par de analizadores, uno basado en el método CYK, descrito en forma algorítmica, y otro basado en el algoritmo Earley, descrito mediante reglas deductivas.

El objetivo principal de la tesis es definir nuevos analizadores para TAG que mejoren el comportamiento práctico de los analizadores tabulares clásicos. Para ello empleamos dos métodos: (1) la extensión del concepto de *Left Corner* de CFG a TAG y su uso para definir esquemas más eficientes para TAG; (2) la construcción de analizadores que trabajen de forma dinámica, comportándose como algoritmos de análisis para TIG cuando analizan árboles TIG y como algoritmos de análisis clásicos para TAG en el resto de situaciones.

Respecto al primer método, las principales aportaciones de la tesis son las siguientes:

- Un nuevo analizador ascendente, al que denominamos **buLC**, obtenido mediante la aplicación de un filtro a **buE**.
- Se define la relación de *Left Corner* en TAG. Se introducen dos analizadores ascendentes con información predictiva, a los que denominamos **pLC** y **LC**, obtenidos mediante la aplicación de un filtro LC al esquema **E**. Además hemos incluido una versión que verifica la propiedad del prefijo válido (VPP).
- Un nuevo analizador, al que denominamos $\mathbf{LC}'$, que disminuye el número de predicciones del esquema **LC**.
- Se llevan a cabo experimentos con un grupo de gramáticas artificiales y un subconjunto de la gramática XTAG, que ponen de manifiesto que los analizadores con filtros LC, tanto en la versión ascendente como predictiva, presentan comportamientos significativamente mejores.
- Demostramos las relaciones formales entre todos los esquemas tipo LC definidos. También probamos las relaciones entre éstos y los esquemas **buE** y **E**, ampliando así la red de analizadores para TAG ya establecida.

Para implementar al segundo método era necesario disponer de un conjunto de analizadores para TIG que empleasen estrategias similares a las de los existentes para TAG. Ante la inexistencia de estos esquemas, nos planteamos como subobjetivo la definición de los mismos, obteniendo los resultados siguientes:

- Un conjunto de nuevos analizadores con diversas estrategias: tres ascendentes ($\mathbf{CYK^i}$, $\mathbf{buE^i}$ y $\mathbf{dVH^i}$) y uno predictivo ($\mathbf{Earley^i}$).
- Se define la relación de *Left Corner* en TIG. Se introducen dos analizadores ascendentes con información predictiva, a los que denominamos $\mathbf{pLC^i}$ y $\mathbf{LC^i}$, obtenidos mediante la aplicación de un filtro LC al esquema $\mathbf{Earley^i}$.
- Demostramos las relaciones formales entre todos los esquemas definidos, creando de esta forma una red de analizadores para TIG similar a la existente para TAG.

De la integración de los esquemas para TAG y TIG con las mismas estrategias obtenemos los siguientes resultados:

- Un conjunto de nuevos analizadores: tres ascendentes ($\mathbf{CYK^{Mix}}$, $\mathbf{buE^{Mix}}$ y $\mathbf{dVH^{Mix}}$) y uno predictivo que no verifica la VPP ($\mathbf{E^{Mix}}$).
- Estudio detallado de la complejidad computacional de todos estos algoritmos con objeto de mostrar en qué casos se producen reducciones de la complejidad.
- Se incluyen experimentos con un grupo de gramáticas artificiales y un subconjunto de la gramática XTAG para comparar los comportamientos prácticos de los esquemas $\mathbf{E^{Mix}}$ y **E**. Comprobándose que el analizador combinado, en general, presenta mejoras apreciables.