Procesamiento del Lenguaje Natural, num. 28 (2002), pp.119-120

Interpretación tabular de autómatas para lenguajes de adjunción de árboles

Miguel A. Alonso Pardo
Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 La Coruña
alonso@udc.es

Resumen: Tesis doctoral en Informática realizada por Miguel A. Alonso Pardo bajo la dirección de los doctores Manuel Vilares Ferro (Universidade da Coruña) y Eric Villemonte de la Clergerie (INRIA, Francia). El acto de defensa de la tesis tuvo lugar el 25 de septiembre de 2000 ante el tribunal formado por los doctores Josep Miró (Universitat de les Illes Balears), José Mira Mira (UNED), Pierre Boullier (INRIA, Francia), Mark-Jan Nederhof (DFKI, Alemania) y Antonio Blanco Ferro (Universidade da Coruña). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad. Se puede obtener más información de la tesis en http://www.dc.fi.udc.es/~alonso/tesis.html.
Palabras clave: Análisis sintáctico, autómatas, gramáticas de adjunción de árboles, gramáticas lineales de índices, programación dinámica.

Abstract: PhD Thesis in Computer Science written by Miguel A. Alonso under the supervision of Dr. Manuel Vilares (Universidade da Coruña, Spain) and Dr. Eric de la Clergerie (INRIA, France). The author was examined in September 25, 2000 by the commitee formed by Dr. Josep Miró (Universitat de les Illes Balears, Spain), Dr. José Mira (UNED, Spain), Dr. Pierre Boullier (INRIA, France), Dr. Mark-Jan Nederhof (DFKI, Germany) and Dr. Antonio Blanco (Universidade da Coruña, Spain). The grade obtained was *Sobresaliente Cum Laude*. Further information is available at http://www.dc.fi.udc.es/~alonso/phd.html.
Keywords: Automata, dynamic programming, parsing, linear indexed grammars, tree adjoining grammars.

1 *Lenguajes de adjunción de árboles*

Las lenguas naturales presentan construcciones sintácticas que no pueden ser descritas mediante gramáticas independientes del contexto. Surge entonces la necesidad de encontrar un formalismo gramatical más adecuado. Puesto que la estructura sintáctica asociada a una frase se representa normalmente como un árbol o, en el caso de frases ambiguas, como un conjunto de árboles, parece natural pensar que un formalismo que manipule árboles y que presente cierta dependencia del contexto debe facilitar la descripción de la sintaxis de las lenguas naturales. En esta dirección las gramáticas de adjunción de árboles (TAG), un formalismo suavemente dependiente del contexto que manipula árboles, se han mostrado adecuadas para la descripción de los fenómenos sintácticos que aparecen en el lenguaje natural.

Los lenguajes generados por las gramáticas de adjunción de árboles constituyen la clase de los lenguajes de adjunción de árboles (TAL), que también son generados por otros formalismos gramaticales suavemente sensibles al contexto como las gramáticas lineales de índices (LIG), más adecuadas al tratamiento computacional que a la descripción lingüística.

2 *Análisis sintáctico*

En la primera parte de la tesis se presenta el problema del análisis sintáctico de los lenguajes de adjunción de árboles. El panorama al comenzar el trabajo que daría lugar a la tesis consistía en un grupo disperso de algoritmos para el análisis sintáctico de gramáticas de adjunción de árboles y apenas un par de algoritmos para el análisis sintáctico de las gramáticas lineales de índices. En esta tesis se ha mostrado que es posible establecer un camino evolutivo continuo en el que se sitúan los algoritmos de análisis sintáctico que incorporan las estrategias de análisis más importantes para ambos formalismos. Los diferen-

ISSN 1135-5948

tes algoritmos se han definido con esquemas de análisis sintáctico, de tal modo que los algoritmos más complejos se derivan a partir de los menos complejos aplicando una secuencia de transformaciones simples. En el caso de las gramáticas lineales de índices el resultado es doblemente interesante, pues si bien se ha esgrimido a su favor su adecuación como formalismo intermedio para el análisis de gramáticas de adjunción de árboles, lo cierto es que numerosas estrategias de análisis para estas últimas no se hallaban incorporadas a ningún algoritmo de análisis sintáctico para LIG. En consecuencia, era necesario sacrificar la estrategia de análisis si se optaba por este enfoque, lo que limitaba enormemente su aplicación práctica. Con el trabajo desarrollado en esta tesis hemos salvado este obstáculo definiendo algoritmos de análisis sintáctico para LIG que incorporan la versión equivalente de las estrategias de análisis más populares para TAG.

3 Modelos de autómata

En la segunda parte de la tesis se definen diferentes modelos de autómata que aceptan exactamente los lenguajes de adjunción de árboles y se proponen técnicas que permiten su ejecución eficiente. La utilización de autómatas para realizar el análisis sintáctico es interesante porque permite optar por un diseño modular al separar el problema de la definición de un algoritmo de análisis sintáctico del problema de la ejecución del mismo, al tiempo que simplifica las pruebas de corrección. Concretamente, hemos estudiado los siguientes modelos de autómata:

- Los autómatas a pila embebidos descendentes (EPDA) y ascendentes (BEPDA), dos extensiones de los autómatas a pila que utilizan como estructura de almacenamiento una pila de pilas. Hemos definido nuevas versiones de estos autómatas en las cuales se simplifica la forma de las transiciones y se elimina el control de estado finito, manteniendo la potencia expresiva.
- La restricción de los autómatas lógicos a pila (RLPDA) para adaptarlos al reconocimiento de las gramáticas lineales de índices, obteniéndose diferentes tipos de autómatas especializados en diversas estrategias de análisis según el conjunto de transiciones permitido.
- Los autómatas lineales de índices, tanto los orientados a la derecha (R–LIA), adecuados para estrategias en las cuales las adjunciones se reconocen de manera ascendente, los orientados a la izquierda (L–LIA), aptos para estrategias de análisis en las que las adjunciones se tratan de forma descendente, como los fuertemente dirigidos (SD–LIA), capaces de incorporar estrategias de análisis en las cuales las adjunciones se tratan de manera ascendente y/o descendente. Este tipo de autómatas utilizan como estructura principal una pila en la que cada uno de los elementos que la componen tiene asociado una pila de índices.
- Los autómatas con dos pilas, una extensión de los autómatas a pila que trabaja con una pila maestra encargada de dirigir el proceso de análisis y una pila auxiliar que restringe las transiciones aplicables en un momento dado. Hemos descrito dos versiones diferentes de este tipo de autómatas, los autómatas con dos pilas fuertemente dirigidos (SD–2SA), aptos para describir estrategias de análisis arbitrarias, y los autómatas con dos pilas ascendentes (BU–2SA), adecuados para describir estrategias de análisis en las cuales las adjunciones se procesan ascendentemente.

Hemos definido esquemas de compilación para todos estos modelos de autómata. Estos esquemas permiten obtener el conjunto de transiciones correspondiente a la implantación de una determinada estrategia de análisis sintáctico para una gramática dada.

Todos los modelos de autómata pueden ser ejecutados en tiempo polinomial con respecto a la longitud de la cadena de entrada mediante la aplicación de técnicas tabulares o de programación dinámica, basadas en la manipulación de representaciones colapsadas de las configuraciones del autómata, denominadas ítems, que se almacenan en una tabla para su posterior reutilización. Con ello se evita la realización de cálculos redundantes.

Finalmente, el análisis conjunto de los diferentes autómatas nos ha permitido determinar sus características comunes y relacionarlos con los analizadores sintácticos tabulares presentados en la primera parte de la tesis.