Procesamiento del Lenguaje Natural, num. 30 (2003), pp.117-118 recibido 1-03-2003; aceptado 20-05-2003

Corrección regional de errores con coste mínimo

Víctor Manuel Darriba Bilbo
Departamento de Informática, Universidad de Vigo
Edificio Politécnico, Campus de As Lagoas, 32004 Orense
darriba@ei.uvigo.es

Resumen: Tesis doctoral en Informática realizada por Víctor Manuel Darriba Bilbao bajo la dirección del doctor Manuel Vilares Ferro de la Universidad de A Coruña. El acto de defensa de la tesis tuvo lugar el 25 de octubre de 2002 ante el tribunal formado por los doctores Roberto Moreno Díaz (Univ. de Las Palmas de Gran Canaria), José Luis Freire Nistal (Univ. de A Coruña), José Gabriel Pereira Lopes (Universidade Nova de Lisboa), Bernard André Dion (Esterel Technologies), Fernando Martín Rubio (Univ. de Murcia). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad.
Palabras clave: Análisis sintáctico, autómatas, programación dinámica, gramáticas independientes del contexto, corrección de errores

Abstract: PhD Thesis in Computer Science written by Víctor Manuel Darriba Bilbao under the supervision of Dr. Manuel Vilares from Universidad de A Coruña(Spain). The author was examined in October 25^{th}, 2002 by the commitee formed by Dr. Roberto Moreno Díaz (Universidad de Las Palmas de Gran Canaria), José Luis Freire Nistal (Universidad de A Coruña), José Gabriel Pereira Lopes (Universidade Nova de Lisboa), Bernard André Dion (Esterel Technologies), Fernando Martín Rubio (Universidad de Murcia). The grade obtained was *Sobresaliente Cum Laude*.
Keywords: Parsing, Automata, dynamic programming, context-free grammars, error repair.

1. Corrección y recuperación de errores

Una cuestión de importancia en el análisis sintáctico, tanto de lenguajes de programación como naturales, es decidir los mecanismos a seguir cuando hay un fallo en la correspondencia del texto con las reglas gramaticales. Los compiladores actuales optan, en general, por devolver un mensaje informando de la localización y el tipo de error y retomando el análisis, bajo la premisa de que el texto debe ser corregido por el usuario y vuelto a ser compilado. Se trata de mecanismos de *recuperación de errores*.

Ahora bien, existe otra posibilidad, más interesante en entornos interactivos, donde no resulta práctica la recompilación continua de textos: las estrategias de *corrección de errores*. Estos algoritmos construyen una o varias modificaciones de la cadena de entrada transformándola *de facto* en una nueva, concordante con la sintaxis del lenguaje. Esto nos permite obtener más rápidamente una entrada analizable sobre la que continuar con el análisis semántico, lo que resulta interesante en el diseño de interfaces hombre/máquina y sistemas de diálogo.

2. Contexto

Tras una primera parte en la que introducimos el problema a tratar y los conceptos básicos relacionados con el análisis sintáctico y la corrección de errores, pasamos en esta segunda parte de la tesis a enfrentarnos a dos cuestiones capitales a la hora de definir nuestro esquema de corrección de errores.

La primera de estas cuestiones es la referente al algortimo de análisis sintáctico que hemos utilizado. A diferencia de la mayoría de los autores de algoritmos de corrección que nos han precedido, nosotros hemos optado por una estrategia de análisis sintáctico no determinista. Esto conlleva la necesidad de gestionar de forma eficiente los diferentes análisis posibles de una entrada ambigua, evitando la repetición de cálculos propia de

ISSN 1135-5948

las estrategias con retroceso. Para ello, utilizamos técnicas de *programación dinámica*, en las que se procura el almacenamiento y compartición de estructuras comunes, lo que redunda en una mejora del rendimiento. En particular, utilizamos el concepto de *entorno dinámico*, para, sobre la base de un autómata con pila determinista, representar todos sus posibles cálculos para una entrada dada de la forma más compacta posible.

La segunda cuestión abordada en la definición de nuestro algortimo se refiere a la determinación del contexto en torno al error en el que se reúne la información necesaria para hallar una corrección. Esta determinación es de gran importancia, puesto que condiciona la aparición de *errores en cascada*, que son aquellos generados por un mal proceso previo de corrección. De hecho, una buena medida de la calidad de un método de corrección es la no generación de este tipo de errores. Existen, en función del contexto en torno al error utilizado, tres grandes familias de algoritmos de corrección:

- *Algoritmos globales.* Trabajan en un contexto global. Esto es, se realizan los mínimos cambios necesarios para corregir las fallas encontradas en la totalidad de la cadena de entrada. La ventaja de estos algoritmos es su calidad, dado que, a la hora de seleccionar una o varias correcciones entre todas las posibles, cuentan con información reunida en el análisis de la totalidad del texto, eliminando de raíz la posibilidad de errores en cascada. El grave inconveniente de estas estrategias es su coste computacional y espacial.
- *Algoritmos locales.* Trabajan en un contexto local fijo, de longitud predeterminada. Cuando se detecta un error, se realizan los mínimos cambios necesarios en la entrada para garantizar que, al menos, puede proseguirse con el análisis de la siguiente palabra. La ventaja de estas soluciones es su bajo coste computacional, puesto que, al contrario que con los métodos globales, no desperdician esfuerzo en áreas de la entrada sin error. Su desventaja es que la información reunida puede ser insuficiente y llevarnos a la elección de malas correcciones y a errores en cascada.
- *Algoritmos regionales.* Son un término medio entre las familias anteriores. Trabajan en un contexto local dinámico. Esto es, se reúne información en una región alrededor del error tan grande como sea necesaria para asegurar la elección de una buena corrección. De este modo, no se explora sistemáticamente la totalidad de la entrada, y no se desperdicia esfuerzo en zonas libres de error; pero tampoco se selecciona la corrección que simplemente nos permita proseguir el análisis de la siguiente palabra de la entrada, de forma que podemos obtener correcciones de calidad comparable a un algoritmo global. Esta es la estrategia elegida para nuestro modelo.

3. Corrección regional con coste mímino

En la tercera parta de la tesis se describe en detalle nuestra propuesta de corrección. Esta se basa en la aplicación de operaciones de edición sobre la cadena de entrada: borrado de una palabra, reemplazamiento por otra, e inserción de una o varias palabras. Por tanto, múltiples análisis son posibles, debido tanto a la consideración de un entorno no determinista como a la aplicación simultánea de diferentes hipótesis de error. Esto nos obliga a fijar un criterio para seleccionar las mejores correcciones. Para ello, asignamos costes individuales para las posibles operaciones de edición. De este modo, el coste de una corrección vendrá dado por la suma de los costes de las inserciones, borrados y reemplazamientos de las que consta, escogiéndose aquellas correcciones con un coste mínimo.

Para determinar la región de corrección utilizamos el concepto de reducción englobando el *punto de error*, la posición de la entrada donde el analizador sintáctico estándar no puede proseguir, y el *punto de detección*, la posición donde se determina que se ha producido un error y se llama al algoritmo de corrección. Esta estrategia regional determina la convergencia asintótica de nuestro algoritmo. En general, se obtienen correcciones de igual calidad que las de los modelos globales con menor coste. Sólo cuando se escogen correcciones de baja calidad que pueden provocar errores en cascada, la región de corrección se va expandiendo para reevaluarlas hasta, en el peor de los casos, abarcar la totalidad de la entrada. Sólo en este caso peor, el coste computacional de nuestro modelo es comparable al de un algoritmo global.