Universidad de Concepción
Facultad de Ingeniería
Departamento de Ingeniería Civil en Informática
Programa Magíster en Ciencias de la Computación

Estudio y Modelamiento de un Shallow Parser de Textos en Lenguaje Natural Utilizando Técnicas de Computación Evolutiva

JUAN EDUARDO MATAMALA PARRA

PROFESOR GUÍA
JOHN ATKINSON ABUTRIDY

Tesis de Grado Presentada a la
ESCUELA DE GRADUADOS
DE LA UNIVERSIDAD DE CONCEPCIÓN

Para optar al Grado de
MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN.

Concepción - Chile
2007

Resumen

El presente trabajo de tesis se enmarca en el área de Procesamiento del Lenguaje Natural (PLN) [15], y más específicamente en el tratamiento de textos en lenguaje natural a nivel sintáctico el cual tiene impacto en aplicaciones tales como Recuperación de Información (Information Retrieval), Extracción de Información (Information Extraction), Minería de textos (Text Mining), etc. Este tipo de aplicaciones se caracterizan por no requerir un tratamiento de textos en profundidad a nivel sintáctico [16], por lo que, podemos contar con un tratamiento superficial, donde no existen estructuras anidadas (superpuestas) lo cual mejora el rendimiento del análisis automático masivo de textos en lenguaje natural. A este tratamiento superficial o intermedio se le denomina *"parsing superficial"* o *"parsing parcial"*.

Actualmente, el tratamiento sintáctico de textos, se realiza a través de variados enfoques, que incluyen los basados en técnicas estadísticas, teoría de la información, y técnicas de aprendizaje automático, entre otros [1, 9, 13]. Estos métodos han logrado buenos resultados en términos de *Precision* y *Recall* [16], sin embargo, presentan algunos problemas o limitaciones que les impide mejorar su rendimiento. En este contexto, esta tesis propone un nuevo enfoque para el tratamiento sintáctico parcial de textos en lenguaje natural basado en técnicas de computación evolutiva, específicamente utilizando algoritmos genéticos. Esto persigue enfrentar algunos de los problemas y limitaciones encontradas en los otros métodos de análisis parcial, como por ejemplo, la precisión y exploración del espacio de búsqueda.

En primer lugar, se realiza un análisis crítico de los principales enfoques actuales de *parsing parcial* [1, 13] con el fin de identificar las limitaciones y ventajas fundamentales en torno a dichos métodos. Luego, se propone un modelo empírico evolutivo de *parsing parcial* basado en algoritmos genéticos, el cual realiza un análisis superficial de textos en lenguaje natural.

En segundo lugar, se describe un prototipo de *parsing parcial* evolutivo, primeramente en base a la estrategia de diseño general del problema de parsing superficial evolutivo, para luego describir específicamente el modelo de chunk parser evolutivo, en base a los elementos o características que definen a un algoritmo genético como lo son: la representación del cromosoma, la población inicial, los operadores genéticos y la función objetivo. Posteriormente, se evalúa la calidad del trabajo del *parser parcial*, para diferentes textos de prueba. Esta evaluación se realiza desde dos puntos de vista: un punto de vista local o interno, y desde un punto vista global o externo, es decir, con respecto a otros métodos de *parsing parcial*.

En la última parte del trabajo, se analiza la veracidad de la hipótesis, y la implicancia de este resultado para el enfoque evolutivo de *parsing parcial* presentado en esta tesis. Finalmente, se dan a conocer las conclusiones globales del trabajo realizado y algunas sugerencias para el desarrollo de un nuevo enfoque evolutivo de *parsing parcial*.