

DESCRIPCIÓN DE UNA NUEVA METODOLOGÍA PARA EL DESARROLLO DE ANALIZADORES SINTÁCTICOS EN EL PROCESAMIENTO DEL LENGUAJE NATURAL (MEDASINTAC)

Gladys Benigni*

RESUMEN

En esta investigación se presenta una metodología para el desarrollo de analizadores sintácticos perteneciente al área de la Inteligencia Artificial, la cual es independiente de la gramática que se seleccione. Es importante destacar que en el Procesamiento del Lenguaje Natural (PLN), específicamente en analizadores, lo que se desea es realizar un análisis a oraciones simples y compuestas con el fin de determinar si la estructura de la misma es o no correcta.

PALABRAS CLAVES: Analizador, lexicón, gramática, metagramática, grafo, reglas gramaticales, procesamiento del lenguaje natural.

ABSTRACT

In this paper, we present a methodology for the development of syntactic analyzers belonging to the Area of Artificial Intelligence, independently from the type of grammar we may choose. It is important to point out that in Natural Language Processing (NLP), specifically in analyzers, what we intend to accomplish is the analysis of simple and compound sentences in order to determine whether their structure is correct or not.

KEY WORDS: Analyzer, Lexicon, Grammar, Metagrammar, Graphs, Grammatical Rules, Natural Language Processing.

INTRODUCCIÓN

La Inteligencia Artificial (IA) es considerada como una de las ramas más importante de la Informática. En las últimas décadas se ha llegado tan lejos en el estudio de la IA y sus aplicaciones, que una de las que quizás más relevancia

y asombro está causando en nuestro tiempo es la idea de pensar que el hombre pueda llegar a establecer una comunicación con el computador, usando como vía para ello su propio lenguaje, ya sea oral o escrito. Esto es lo que se conoce como Procesamiento del Lenguaje Natural (PLN).

En el presente trabajo se presenta el desarrollo de una metodología que permite generar analizadores sintácticos en nuestro idioma, basándonos en los estudios y propuestas de diferentes autores.

MÉTODOS

Se utilizaron los pasos para el análisis morfológico, semántico y sintáctico del Procesamiento del Lenguaje Natural propuesto por Rich y Knight (1994). Para los ejemplos se hizo referencia a la gramática transformacional y de dos niveles indicada por Krulee (1991).

DESARROLLO DE LA METODOLOGÍA

La siguiente metodología ha sido dividida en tres modelos o fases: Análisis, Diseño e Implantación, basado en el estudio del análisis morfológico y sintáctico de oraciones en nuestro propio idioma (El Castellano).

1.- Modelo de Análisis

En esta etapa, debe realizarse una minuciosa revisión de bibliografía en lingüística, a objeto de seleccionar la gramática requerida para el desarrollo del analizador sintáctico. Este estudio detallado es importante para evitar problemas de concordancia, diferencia de significado entre expresiones, etc. La gramática es la representación declarativa de los hechos sintácticos sobre el lenguaje; también se define como un conjunto de reglas o principios que dicen cómo hablar y entender un idioma.

El estudio de la gramática, se lleva a cabo a través del análisis morfológico, sintáctico, semántico y pragmático si así lo requiere el sistema. El análisis morfológico debe realizarse para asignar a cada una de las palabras constituyentes de la oración su respectiva categoría sintáctica, esto es, sustantivo, verbo, adverbio, entre otros.

*Departamento de Informática. Universidad de Oriente. Núcleo de Nueva Esparta.

Recibido: Octubre 1998. Aprobado: Mayo 1999.

Una vez asignadas las categorías a cada palabra que constituye la oración, se realiza el análisis sintáctico, el cual construye una estructura completa de la oración (generalmente en forma de árbol), dependiendo de las reglas gramaticales. En el análisis semántico se lleva a cabo una correspondencia de significado entre todas las palabras que constituyen la oración, esto es, género, número y persona.

2.- Modelo de Diseño

Para ingresar a esta fase se seleccionó el tipo de gramática a ser utilizada para la construcción del analizador. Al desarrollar sistemas de este tipo, debe integrarse una serie de componentes que permitan analizar una oración y emitir un resultado. Para esto, debe diseñarse el *analizador*, considerado como el corazón del sistema en el Procesamiento del Lenguaje Natural.

En la Figura 1, puede observarse el diseño general de un sistema que analiza oraciones en su nivel morfológico y sintáctico.

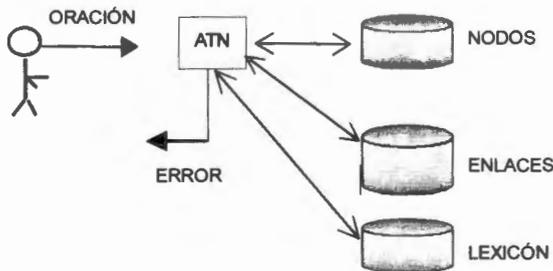


Fig.1. Diseño general del sistema

En la Figura 1, la gramática ATN (Red de Transiciones Aumentadas) representa el analizador; éste se encarga de analizar la oración de entrada, palabra por palabra, para asignarle la categoría correspondiente, tomando para esto aquellas que se encuentran en el lexicón o diccionario. Rich (1994) refiere que “una red de transiciones aumentada es un procedimiento de análisis sintáctico descendente que permite incorporar al sistema varias clases de conocimiento de forma que pueda funcionar eficientemente”. El lexicón contiene las palabras que serán utilizadas por el analizador, constituyéndose de esta forma el diccionario del sistema.

El analizador estará completo, una vez se haya diseñado la gramática seleccionada, el lexicón, la metagramática y la red de transición de estados.

2.1.- Diseño de la Gramática

Una vez seleccionada la gramática, la misma debe representarse a través de Reglas Gramaticales o Reglas de Producción. A continuación, ejemplificaremos las reglas gramaticales basándonos en la gramática transformacional (reconocimiento de dos niveles estructurales: estructura superficial y estructura profunda; la primera representa la expresión tal cual como se dice y escribe, la segunda representa de una manera lógica lo que realmente significa la expresión) y de dos niveles (esta gramática usa dos conjuntos de reglas, una representa la gramática y la otra la metagramática).

- 1.- $O \rightarrow SN SV$
- 2.- $SN \rightarrow$ Determinante Sustantivo (Adjetivo) (SP)
- 3.- $SN \rightarrow$ Pronombre (SP)
- 4.- $SV \rightarrow$ Auxiliar Verbo (Adverbio) (SN | SP)
- 5.- $SP \rightarrow$ Preposición SN (SP)

En donde:

O = Representa la oración.

SN = Sintagma Nominal.

SP = Sintagma Preposicional.

SV = Sintagma Verbal.

Los símbolos que se encuentran dentro de paréntesis, pueden ser omitidos, y la barra vertical que separa dos o más símbolos indica que se debe elegir uno y solo uno de los símbolos que separa.

En las Figuras 2.a y 2.b, se reflejan algunas de las reglas gramaticales descritas anteriormente, ampliando su contexto.

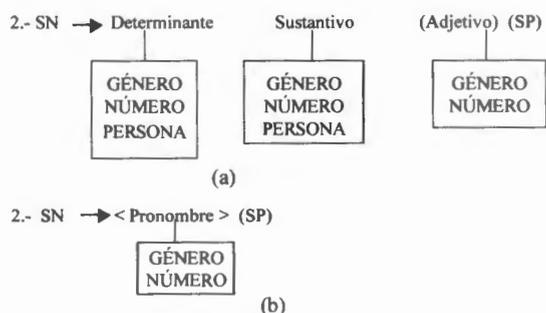


Fig. 2. Ampliación de dos Reglas Gramaticales.

Es importante mostrar el diseño de una red de transición de estados, utilizando las reglas gramaticales establecidas. La figura 4 representa el diseño para una gramática transformacional y de dos niveles (gramática que ha sido utilizada en todo el trabajo para ejemplificar).

A este punto, es necesario definir la metagramática, tal como se muestra a continuación.

- 1.- Número → Singular | Plural
- 2.- Género → Masculino | Femenino
- 3.- Persona → 1 | 2 | 3

2.2.- Diseño del Lexicón

El lexicón o diccionario es un conjunto de palabras que contiene las características esenciales de las mismas, cómo la categoría léxica a la cual representa y cómo puede insertarse en la estructura profunda por medio de una regla individual. Lyons (1981), plantea que la teoría estándar de la gramática transformacional, indica que la oración tiene dos niveles distintos de estructura sintáctica conectados mediante reglas de una clase especial llamada transformacionales. Estos niveles son la estructura profunda y la estructura superficial. La estructura profunda está más íntimamente conectada con el significado de la oración de lo que está la estructura superficial. Por otro lado, la estructura superficial, está más íntimamente conectada con la forma en que se pronuncia la oración.

Para diseñar el lexicón, es necesario representar una tabla o archivo donde se indique para cada palabra su categoría, persona, número, género y forma. En la Figura 3, se observa un ejemplo de cómo representar esta tabla o archivo.

PALABRA	CATEGORIA	PERSONA	NUMERO	GÉNERO	FORMA
baila	Verbo	3ra.	S	-	Presente
a	Preposición	-	-	-	-
tomaré	Verbo	1ra.	S	-	Futuro
abierta	Adjetivo	-	S	F	-

Fig. 3 Representación del lexicón.

Retomando la Figura 3, describiremos las categorías léxicas representadas en la misma.

Palabra: Es la palabra del lenguaje o idioma seleccionado.

Categoría: Categoría correspondiente a cada palabra, tal como verbo, adjetivo, sustantivo, preposición, etc.

Persona: Persona a la que pertenece cada palabra, esto es, primera, segunda o tercera.

Número: Indica si la palabra se expresa en una o varias personas.

Género: Indica si la palabra es femenina o masculina.

Forma: Representa el estado del verbo, esto es, presente, pasado, futuro, entre otros.

2.3.- Diseño de la Red de Transición de Estados

Una vez definida la gramática, se construye la red de transición de estados, tomando en cuenta las reglas gramaticales ya definidas. La manera idónea de diseñar la red, es a través de un grafo dirigido $G = (N, E)$ en donde N representa los nodos y E los enlaces respectivos. Los enlaces (E), deben etiquetarse por categoría, esto es pronombre, determinante, sustantivo, adjetivo, entre otros.

Los nodos, simplificando, pueden clasificarse como nodo inicial, nodo intermedio y nodo final. El nodo inicial puede ser una oración, sintagma nominal, sintagma verbal, entre otros. Los nodos intermedios, indican la descomposición de una regla gramatical; si se trata de un sintagma nominal, tendríamos por ejemplo la siguiente descomposición:

$$SN \rightarrow \langle \text{Determinante} \rangle \langle \text{Sustantivo} \rangle (SP)$$

El nodo final, indica un destino que conduce a la veracidad o falsedad de una regla gramatical. En este caso es retornar al usuario un mensaje de que la oración que introdujo es correcta o no.

Es importante mostrar el diseño de una red de transición de estados, utilizando las reglas gramaticales establecidas. La figura 4 representa el diseño para una gramática transformacional y de dos niveles (gramática que ha sido utilizada en todo el trabajo para ejemplificar).

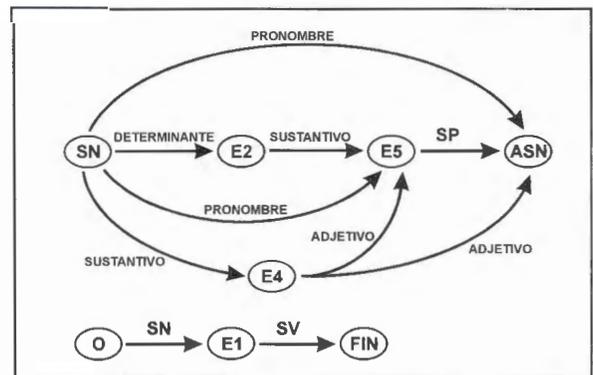


Fig. 4. Red de Transición de Estados.

En la Figura 4, se puede observar que los nodos iniciales para las dos reglas gramaticales propuestas corresponden a una oración (O) y a un sintagma nominal (SN).

3.- Modelo de Implantación

Una vez elaborada las fases de análisis y diseño, se procede a seleccionar el hardware y software necesario para satisfacer las necesidades del sistema. Una vez seleccionados los requerimientos computacionales se realizaría la programación del mismo.

Posteriormente, deben efectuarse las pruebas necesarias al sistema, para verificar el funcionamiento del mismo. Estas pruebas, consisten en introducir oraciones para determinar si su estructura es correcta, de lo contrario debe indicar el mensaje de error correspondiente.

Finalmente, se elabora el manual de usuario, a fin de describir el uso del sistema, de manera tal que cualquier usuario no identificado con el mismo pueda interactuar fácilmente.

CONCLUSIÓN

La importancia de este trabajo es haber introducido una nueva herramienta para desarrollar sistemas en el área del Procesamiento del Lenguaje Natural, particularmente en lo que se refiere a los analizadores sintácticos. La misma facilita el desarrollo paso a paso evitando el tener que ajustar o combinar métodos, que no garantizan el desarrollo adecuado de sistemas de este tipo.

AGRADECIMIENTO

Al Consejo de Investigación de la Universidad de Oriente y particularmente al Prof. Luis León. De igual forma a los bachilleres Efrén Fuentes y Mónica Rodríguez por enfrentarnos al reto del Procesamiento del Lenguaje Natural en su Trabajo de Grado. En el área de lingüística a la Prof. Florelia Páez.

BIBLIOGRAFÍA

- KRULLEE, G. 1991. Computer Processing of Natural Language. Ediciones Prentice-Hall Internacional. Englewood Cliffs. New Jersey. pp. 45 – 108.
- LEVINE, R. 1990. AI and Expert Systems. A Comprehensive Guide, C Language. Second Edition. McGraw-Hill, Inc. pp. 25 – 29.
- LYONS, J. 1981. Lenguaje, Significado y Contexto. Ediciones PAIDÓS Ibérica, S.A. Barcelona-Buenos Aires-México. pp. 152 – 161.
- PATTERSON, D. 1990. Introduction to Artificial Intelligence and Expert Systems. Prentice-Hall International Editions. pp. 229 – 254.
- RICH, E. Y KNIGHT, K. 1994. Inteligencia Artificial. Segunda Edición. Editorial McGraw – Hill / Interamericana de España, S.A. pp. 418 – 421.
- SHILDT, H. 1990. Utilización de C en Inteligencia Artificial. Osborne/McGraw-Hill Interamericana de España, S. A. pp. 99 – 102.