CARLOS A. FERNÁNDEZ CAIRÓ*
JORGE A. RÍOS GARCÍA*
YUNIOR RAMÍREZ CRUZ*, **
DATYS: Desarrollo de Aplicaciones, Tecnologías y Sistemas*
Centro de Estudios de Reconocimiento de Patrones y Minería de Datos, Universidad de Oriente**
Santiago de Cuba, Cuba
{carlos.fernandez, jorge.rios}@sc.datys.co.cu, yunior@cerpamid.co.cu

Aproximaciones para el análisis sintáctico del español

1 Introducción

Una de las dificultades del Procesamiento del Lenguaje Natural (PLN) de forma automática mediante un ordenador es la ambigüedad. El Lenguaje Natural es inherentemente ambiguo a diferentes niveles, por ejemplo, *léxico*, *referencial*, *estructural* y *pragmático*. En la actualidad, los problemas de ambigüedad relacionados con la sintaxis del lenguaje son ampliamente tratados. El problema central del Análisis Sintáctico (AS) en el PLN es la traducción de entradas en lenguaje natural a una representación interna sin ambigüedad estructural, como los árboles de análisis. Como fase del PLN, el AS realiza un análisis de la estructura sintáctica de una frase a través de una gramática propia del lenguaje. Éste puede ser clasificado en *AS de constituyentes*, basado en las relaciones de inclusión de sintagmas y unidades léxicas, o *AS de dependencias*, que se caracteriza por establecer relaciones de dependencia entre las unidades léxicas [2].

El AS de constituyentes se divide en superficial o profundo. En el *AS de constituyentes superficial* sólo se detectan un subconjunto de las estructuras presentes en una oración, mientras que en el *AS de constituyentes profundo* se obtiene el árbol sintáctico completo.

En el presente trabajo se describe de manera detallada una extensión del conjunto de reglas del analizador sintáctico de la herramienta FreeLing [3], relacionadas con la estructura sintáctica de una oración y las relaciones de dependencia entre algunos pares de sintagmas. Tal extensión consiste en una ampliación del conjunto de reglas de la herramienta FreeLing, mediante las cuales se intenta completar algunas relaciones de dependencia entre las unidades léxicas. Se presenta además la estrategia seguida para la construcción de un Analizador Sintáctico de dependencias con el uso de técnicas de aprendizaje automático que se basa en el método MST de McDonald [1], en el cual se obtiene un árbol de dependencias a partir de un grafo dirigido que contiene las relaciones de dependencia existentes entre las unidades léxicas correspondientes.

Por último, se presenta un método para obtener un árbol de constituyentes profundo realizando movimientos en el árbol superficial de acuerdo con la información del árbol de dependencias.

2 Aproximaciones presentadas

En esta sección se detallan las dos aproximaciones al análisis sintáctico que se proponen en este trabajo. Primeramente, se presenta la aproximación basada en el conocimiento, específicamente basada en reglas, y luego se describe el enfoque basado en técnicas del aprendizaje automático.

2.1 Aproximación basada en el conocimiento

Aunque existen muchos trabajos para el AS del Inglés; no sucede lo mismo para el español, entre los trabajos realizados para el español se encuentra el FreeLing [3], herramienta de código abierto que brinda entre sus servicios el análisis sintáctico basado en conocimiento. Esta herramienta realiza el AS de constituyentes superficial y el AS de dependencias.

Para realizar el AS de constituyentes superficial el FreeLing utiliza como punto de partida un conjunto de oraciones cuyas palabras poseen rasgos de interés lingüístico como lema, etiqueta morfosintáctica y otros. Luego, aplicando a cada una de estas oraciones un conjunto de reglas (*gramática del lenguaje*), son determinados los constituyentes (*categorías gramaticales*) presentes en la oración. Este proceso da como resultado un árbol sintáctico donde los nodos hojas son las palabras presentes en la oración y el resto de los nodos los sintagmas encontrados en la oración.

En el otro caso, se utiliza como punto de partida, al igual que en el primer caso, el mismo conjunto de oraciones pero esta vez cada oración cuenta con el árbol obtenido en el AS de constituyentes superficial de la misma. Luego, a partir del las categorías gramaticales presentes en la oración y un conjunto de reglas de dependencia entre categorías gramaticales y etiquetas (*tipo de dependencia*) para cada dependencia determinada, se encuentran las relaciones de dependencia existente entre las unidades léxicas presentes en la oración, se etiquetan las mismas y se obtiene un árbol construido a partir de las reglas aplicadas.

2.1.1 Extensión del conjunto de reglas del FreeLing

La principal motivación para extender el conjunto de reglas del analizador sintáctico del FreeLing fue la presencia de algunas limitaciones en su funcionamiento, fundamentalmente en cuanto a la manera de manipular algunos tipos de sintagmas presentes en una oración, las relaciones de composición entre los sintagmas y las unidades léxicas. De manera análoga, la extensión del conjunto de reglas del AS de dependencias se realizó para detectar la presencia de nuevas relaciones de dependencia entre algunos pares de sintagmas.

Algunos casos como la composición de *sintagmas verbales,* de *sintagmas nominales* o *sintagmas preposicionales* fueron reescritos o añadidos para el funcionamiento del Analizador Sintáctico de Constituyentes Superficial.

Por su parte, en el Análisis Sintáctico de Dependencias, situaciones como la relación de dependencia entre una oración subordinada y la principal y relaciones de dependencia no previstas formaron parte de la extensión de este análisis.

Relaciones de composición de sintagmas

Se introdujeron nuevas relaciones de composición de sintagmas entre sintagmas nominales y sintagmas preposicionales para el análisis de constituyentes. Otras, como en el caso de los sintagmas verbales compuestos por un verbo y un verbo infinitivo (por ejemplo, *puedo ayudar*) se reescribieron de modo que la función sustantiva del infinitivo sea tenida en cuenta y se considere la forma verbal conjugada como núcleo. Estas extensiones se basaron principalmente en consideraciones lingüísticas.

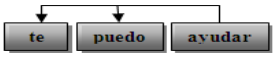


Fig. 1. Composición de un sintagma verbal realizada por el FreeLing

Fig. 2. Nueva composición del sintagma verbal

Relaciones de dependencia

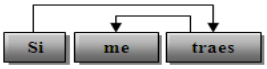


Fig. 3. Oración subordinada

+

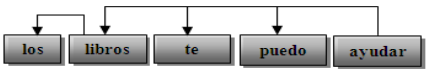


Fig. 4. Sintagma verbal representa la oración principal

=

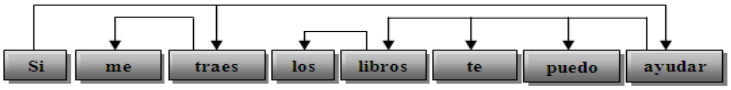


Fig. 5. Árbol de Dependencia del FL. La oración principal depende de la oración subordinada

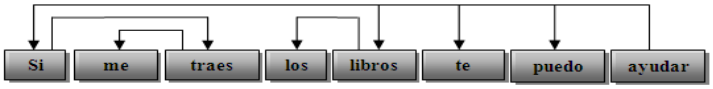


Fig. 6. Árbol de Dependencia de la extensión del FL. La oración subordinada depende de la oración principal

Para la extensión de las relaciones de dependencia con las que cuenta el FreeLing, se tomaron en cuenta casos como relaciones de dependencia no previstas y la relación de dependencia entre una oración subordinada y su oración principal, la cual no se manipula bien en todos los casos. Desde la Figura 3 a la Figura 6 se muestra el resultado del proceso de búsqueda de la relación de dependencia entre una oración subordinada y su oración principal. La Figura 5 muestra la relación de dependencia determinada de manera incorrecta por el FreeLing, mientras que la 6 muestra la relación de dependencia lograda en la extensión del AS de dependencias, la cual consideramos correcta.

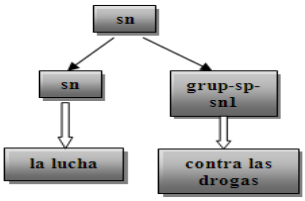
2.1.2 Análisis Sintáctico de constituyentes profundo

A esta aproximación se añadió el AS de constituyentes profundo. El árbol correspondiente a este análisis se obtiene a partir del árbol sintáctico superficial y el árbol de dependencias. El proceso se basa en tener en cuenta la información sobre las dependencias existente en el árbol de dependencias entre las categorías sintácticas y realizar los movimientos pertinentes en el árbol superficial para reflejar dichas dependencias entre los constituyentes y así obtener el árbol de constituyentes profundo.

En las figuras que se muestran a continuación, las flechas negras indican dependencia directa y las flechas blancas dependencia por transitividad. En este trabajo, para la construcción del árbol de constituyentes profundo se han tenido en cuenta tres tipos de dependencias:

1. Sintagmas preposicionales subordinados a un sintagma nominal

   Con la oración "*El Gobierno dará prioridad a la lucha contra las drogas*" (*oración 1*) se mostrará el movimiento a realizar en el árbol superficial.



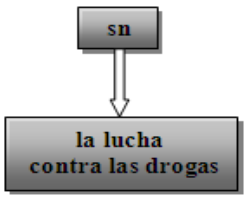


Fig. 7. Fragmento del Árbol Sintáctico de Constituyentes Superficial correspondiente a la oración 1

Fig. 8. El sintagma preposicional se subordina al sintagma nominal

En la figuras 7 y 8 se muestran los constituyentes encontrados en la oración 1 y el movimiento del sintagma preposicional *contra las drogas* que es parte del sintagma nominal *la lucha* para colocarlo como hijo de éste en el árbol, quedando *la lucha contra las drogas* como una rama del árbol profundo.

2. Oraciones coordinadas

   Con la oración "*Jorge trajo un perrito y se lo regaló a la prima de su vecina*" (*oración 2*) se mostrará el movimiento realizado en el árbol superficial para reflejar las dependencias. En la *figura 9* se muestra el árbol profundo obtenido teniendo en cuenta la relación de dependencia de una oración formada por dos oraciones coordinadas.

3. Oraciones Subordinadas

   Con la oración "*Me pidieron que trajera un libro de inglés*." (*oración 3*) se mostrará el movimiento a realizar en el árbol superficial para reflejar las dependencias. En la *figura 10* se muestra la relación de dependencia entre una oración subordinada y la oración principal.

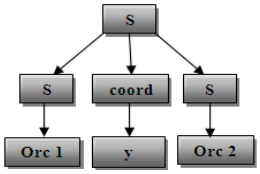


Fig. 9. Árbol profundo donde se delimitan las oraciones 1 y 2 que forman a la oración principal

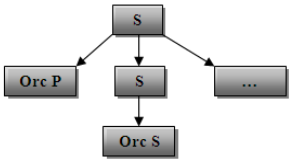


Fig. 10. Árbol profundo donde se delimita la oración subordinada con respecto a la principal

2.2 Utilización de técnicas de aprendizaje

Preliminares:

Nuestro método se basa en el trabajo de McDonald [1]. Denotemos por $x=x_1,\ldots,x_n$ a la oración de entrada al análisis, y sea y su árbol de dependencias que estará representado por un conjunto de aristas entre pares de elementos (x_i, x_j) de la oración. Se dice que $(i, j) \in y$ si existe en y la dependencia $x_i \rightarrow x_j$ y $s(i, j) = w$ es su peso.

Groso modo, en esta aproximación primeramente se construye un grafo dirigido $G = (V, E)$ donde $V = \{x_1,\ldots, x_n\}$ y $E \subseteq [1{:}\, n] \times [1{:}\, n]$. Cada arista en E es pesada con un valor real w, el cual es este trabajo se obtiene a partir de una red neuronal Perceptron con una capa oculta. Luego, se define el árbol de dependencias y como el árbol se soporte máximo de G, $y \subseteq E$, que maximiza el valor $\sum_{(i,j)\in y} s(i,j)$ y de forma tal que todos los vértices de V aparecen en y.

Específicamente, el grafo de cada oración x, G_x, está formado por las palabras de x y el símbolo *root* como vértices. Se coloca además una arista del símbolo *root* a cada palabra.

El árbol de soporte máximo se halla mediante el algoritmo ChLE [4].

Algoritmo ChLE

El algoritmo que se muestra a continuación es el que obtiene el MST.

ChLE (G, s)

Grafo $G = (V, E)$

Función de peso de las aristas $s{:}\ E \rightarrow \mathcal{R}$

1. $M=\{(x^*,x):x\epsilon V,\ x^*=argmax_{x'}\ s(x',x)\}$
2. $G_M=(V,M)$
3. Si G_M no tiene ciclos, entonces es un MST: retornar G_M
4. sino: encontrar un ciclo C en G_M
5. $G_C=Contract\ (G_C,\ s)$
6. $Y=ChLE\ (G_C,\ s)$
7. Encontrar vértice $x\epsilon C$ tal que $(x',x)\epsilon Y,\ (x'',\ x)\epsilon C$

8. Retornar Y $\cup$ C-{(x", x)}

Contract (G=(V, E), C, s)

1. G_C es el subgrafo de G excluyendo los nodos en C
2. Adicionar nodo c a G_C representando el ciclo C
3. Desde $x \in V - C : \exists_{x' \in C} (x', x) \in E$. Adicionar arista (c, x) a G_C con $s(c, x) = max_{x' \in C} s(x', x)$
4. Desde $x \in V - C : \exists_{x' \in C} (x, x') \in E$. Adicionar arista (x, c) a G_C con $s(x, c) = max_{x' \in C}[s(x, x') - s(a(x'), x') + s(c)]$ donde $a(v)$ es el predecesor de v en C y Retorna G_C.
5. $s(C) = \sum_{v \in C} s(a(v), v)$

Obtención de los pesos

Como se mencionó antes, para obtener los pesos se usó una red neuronal (Perceptrón simple), la cual consta de una única capa oculta y una neurona de salida [5].

Para el cálculo de la entrada total se usa la siguiente ecuación: $s_i = \sum_j w_{ji} u_j + bias_i$ donde u_j es la entrada x_j o la salida de una neurona y_i. $u_0 = 1$. El $bias$ es la predisposición de la neurona a activarse, o valor inicial añadido al cálculo de la activación. Normalmente el $bias$ se trata como un peso más en cada neurona.

El cálculo del nivel de activación se realiza mediante el modelo continuo: $y_i = \frac{1}{1+e^{-s_i}}$, que es el más utilizado en la literatura [5]. Cada y_i representa una función sigmoidea (corta al eje de las *y* en *0.5* y está acotada en *0* y *1*). El algoritmo de entrenamiento usado es el de retro-propagación (backpropagation) que es mostrado a continuación.

Backpropagation

1. Inicializar los pesos con valores aleatorios pequeños entre *-0.1* y *0.1*
2. Presentar un patrón *(X, D)* donde $D_i = 1$ si pertenece a la clase *i*, y *0* en otro caso.
3. Calcular las salidas reales de todas las neuronas según el modelo continuo.
4. Calcular el error de todas las neuronas. Si es de salida: $E_j = (y_j - D_j) \cdot y_j \cdot (1 - y_j)$. Si es oculta: $E_j = y_j \cdot \sum_k z_k \cdot w_{jk}$ donde k recorre todas las neuronas conectadas a la j en la capa siguiente y z_k son los errores de ellas.
5. Ajustar los pesos: $w_{ij}(n+1) = w_{ij}(n) + \alpha E_j y_j + \mu\left(w_{ij}(n) - w_{ij}(n-1)\right)$, donde α es la velocidad de aprendizaje y μ es llamado momento.

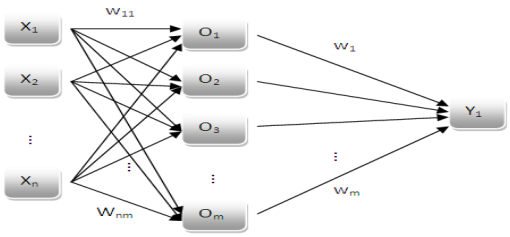


Fig. 11. Esquema del perceptrón simple utilizado

3. Conclusiones

En este trabajo se presentaron dos aproximaciones para el análisis sintáctico en español, las cuales se encuentran actualmente en desarrollo. Una de ellas persigue extender el conjunto de reglas de la herramienta FreeLing, para agregar algunas relaciones de dependencia entre las unidades léxicas que en la herramienta original no se manipulan. Aún cuando los cambios introducidos son correctos desde el punto de vista lingüístico, aún es necesario realizar un análisis experimental adecuado que demuestre la validez de los mismos en corpus de diferentes dominios.

La segunda aproximación utiliza una red neuronal para calcular los pesos en el algoritmo basado en el árbol de soporte máximo de McDonald. Al igual que para el primer caso, aún es necesario un análisis experimental adecuado para poder llegar a conclusiones acerca de la validez de este método.

Adicionalmente, hemos presentado un método mediante el cual se obtiene el árbol de constituyentes profundo realizando movimientos en el árbol superficial guiados por el árbol de dependencias.

Como trabajo futuro, además de la necesaria evaluación experimental de ambos métodos, se pretende explorar estrategias de combinación de los mismos. Además, los métodos presentados deben ser integrados dentro de la suite de herramientas para el Procesamiento del Lenguaje Natural para la cual están siendo desarrollados.

Referencias bibliográficas

[1] McDonald, Ryan. Online Large-Marging of dependency parsers. Proceedings of the 43rd Annual Meeting of the ACL, page 91-98. Junio 2005.

[2] Pedro J. Moriano, Luis Romero, Alfonso Muñoz. JBeaver: Análisis y Generación de Analizadores de Dependencia.

[3] FreeLing: http://garraf.epsevg.upc.es/freeling

[4] McDonald, Ryan. Non-projective dependency parsing using spanning tree algorithms. Human Language Technologies and Empirical Methods in Natural Language Processing (HLT-EMNLP). 2005

[5] Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press. ISBN 0-19-853849-9 (hardback) or ISBN 0-19-853864-2