Herramienta para el etiquetado léxico y análisis sintáctico de textos orientado a la construcción de corpus supervisados

Xavier Ribera, Antonio Molina, Ferran Pla
Universitat Politècnica de València
Departament de Sistemes Informàtics i Computació
Camí de Vera s/n, 46020 València
xarimas@fiv.upv.es,{fpla,amolina}@dsic.upv.es

Resumen En este artículo se describe una aplicación gráfica útil para el análisis sintáctico y la supervisión de textos. Esta aplicación integra diversas herramientas para el etiquetado léxico y el análisis sintáctico parcial de las oraciones de entrada. El usuario puede corregir tanto el árbol de análisis como el etiquetado léxico producido por métodos automáticos, hasta conseguir el análisis completo de la oración. Esto permite reducir el tiempo de supervisión de un corpus, que podrá ser utilizado posteriormente para el entrenamiento de etiquetadores léxicos o analizadores sintácticos que utilizan técnicas de aprendizaje basadas en corpus.

1 *Introducción*

Durante los últimos años se han desarrollado multitud de técnicas de análisis en los distintos niveles de procesamiento del lenguaje natural (léxico, sintáctico y semántico) que se basan en métodos de aprendizaje. Dichas técnicas necesitan de un corpus de entrenamiento anotado con la información que se debe aprender (etiquetas léxicas, parentizado sintáctico, roles sintácticos, etiquetas semánticas, etc). La construcción de corpora anotados supervisados es una tarea costosa debido a que requiere la intervención de un experto para que revise el resultado producido por los distintos desambigüadores automáticos. Esta tarea de supervisión es aconsejable para que los métodos de aprendizaje automático utilicen en la medida de lo posible datos libres de errores. Para ello, sería deseable facilitar al usuario experto (normalmente lingüístas) las tareas de supervisión mediante entornos gráficos amigables que disminuyan el coste de corrección del corpus.

Este artículo describe una herramienta gráfica que facilita la supervisión de los corpora de entrenamiento a partir del resultado proporcionado por un proceso previo de análisis. Este proceso consta de tres fases que se desarrollan de forma secuencial: análisis morfológico, etiquetado léxico y análisis sintáctico parcial. La aplicación permite que el usuario pueda, de manera interactiva, corregir la salida proporcionada por cualquiera de estas fases antes de continuar con la siguiente.

En el apartado 2 se describen las herramientas integradas en la aplicación: 1) El analizador morfológico [Carmona98] que segmenta el texto en unidades léxicas y proporciona para cada una de ellas todas las posibles etiquetas léxicas. 2) El etiquetador léxico [Pla99] basado en modelos regulares estocásticos: modelos de n-gramas o modelos regulares inferidos mediante el algoritmo *ECGI* [Prieto92]. 3) El analizador sintáctico parcial basado en expresiones regulares [Molina99]. 4) El análisis integrado que permite llevar a cabo el etiquetado léxico y el análisis sintáctico superficial del texto en una sola fase (utilizando también modelos regulares) [Pla00].

La funcionalidad de la aplicación se describe en el apartado 3. La herramienta permite al usuario: trabajar con las etiquetas léxicas definidas en *PAROLE* [Marti98], así como con las etiquetas utilizadas en el proyecto *Penn TreeBank* [Marcus93]; editar estas etiquetas o crear su propio conjunto de etiquetas; definir las reglas gramaticales que se utilizarán en el análisis parcial usando las etiquetas léxicas como terminales; corregir el resultado del etiquetado léxico sobre el texto analizado; corregir el análisis parcial sobre el árbol de análisis; completar el análisis resolviendo manualmente la ambigüedad estructural. Además, permite evaluar las prestaciones de los distintos analizadores.

Esta herramienta se ha desarrollado en un entorno Linux utilizando el lenguaje *tcl-tk v.8.0.5-30.*

2 *Descripción de los niveles de análisis*

2.1 Análisis morfológico

El análisis morfológico, para textos no restringidos en castellano, se realiza mediante el analizador morfológico *MACO* [Carmona98]. Esta herramienta segmenta los textos de entrada en unidades léxicas o 'tokens', identificando signos de puntuación, abreviaciones, números, nombres propios, agrupaciones léxicas tales como locuciones adverbiales, fechas, etc. Como salida proporciona para cada 'token' detectado, todas sus posibles etiquetas léxicas. Si el 'token' es desconocido, se considera que puede pertenecer a cualquier categoría abierta (nombre común, adjetivo, verbo, etc).

En el caso de que se trabaje con textos en inglés, aunque la aplicación no dispone de un analizador de estas características, se ha simulado construyendo un diccionario con la información obtenida del corpus etiquetado *Penn TreeBank* [Marcus93]. Esta aproximación, aunque permite trabajar con textos del citado corpus, se debería extender para textos no restringidos utilizando algún analizador morfológico que incluya el tratamiento de las palabras desconocidas.

2.2 Etiquetado léxico

Este sistema contempla, aunque no están integrados en la interfaz gráfica de la aplicación, una serie de modulos que llevan a cabo el proceso de entrenamiento de los diferentes modelos involucrados en la tarea de desambiguación léxica, Modelo de Lenguaje (ML) y Modelo Léxico, utilizando diferentes aproximaciones.

Los etiquetadores que incorpora están basados en modelos regulares estadísticos: bigramas y autómatas de estados finitos aprendidos automáticamente utilizando técnicas de Inferencia Gramatical (*ECGI*) [Pla98]. Los Modelos de Lenguaje se infieren a partir de corpus etiquetados léxicamente. A estos modelos se les aplican técnicas de suavizado para garantizar la completa cobertura del lenguaje: en nuestro caso los modelos basados en bigramas se han suavizado mediante técnicas de 'back-off' y para los modelos *ECGI* se han desarrollado distintos suavizados utilizando la técnica de 'back-off' y la de interpolación lineal [Pla99]. El Modelo Léxico se estima, igualmente a partir de texto etiquetado, calculando las frecuencias de aparición de las palabras, de las categorías y de cada palabra en cada categoría. Estos datos son necesarios para estimar las probabilidades léxicas, las cuales también han sido convenientemente suavizadas.

Como el formalismo utilizado para representar los diferentes modelos del lenguaje es el mismo (máquinas de estados finitos), el proceso de etiquetado también es homogéneo. Dada una frase de entrada, el problema consiste en encontrar en el autómata la secuencia de estados de mayor probabilidad, para lo que se utiliza el algoritmo de Viterbi [Viterbi67]. Una vez obtenida la secuencia de estados óptima, como cada estado tiene asociada una única categoría léxica, se dispone de la mejor secuencia de categorías para la secuencia de palabras ('tokens') de entrada y éste es su etiquetado.

2.3 Análisis sintáctico parcial

El analizador sintáctico incorporado en la herramienta (APOLN, Analizador Parcial de Oraciones en Lenguaje Natural [Molina99]), permite realizar el análisis parcial de oraciones escritas en lenguaje natural no restringido. Está construido utilizando técnicas de máquinas de estados finitos. Su funcionamiento es incremental, es decir, analiza las oraciones aplicando una secuencia de pasos o niveles de procesamiento, de manera que la entrada a un determinado nivel es la salida del nivel inmediatamente anterior. En cada uno de estos niveles se reconoce un conjunto de estructuras sintáctica o patrones que se describen mediante definiciones regulares. Estos patrones se definen utilizando como alfabeto el conjunto de etiquetas léxicas y de patrones. Además son patrones no recursivos, es decir, en la definición de un nivel determinado sólo pueden aparecer patrones definidos en niveles previos. APOLN toma como entrada una oración previamente etiquetada léxicamente. En cada nivel se parentiza la entrada agrupando las secuencias de símbolos reconocidas por alguno de los patrones definidos en ese nivel y esto constituye la entrada para el siguiente nivel.

Un ejemplo de definiciones de patrones sencillas para el castellano, utilizando las etiquetas léxicas *PAROLE*, se muestra a continuación:

Nivel 1 // núcleos nominales y verbales
NSN –> (NC | NP)+
NSV –> (VM | VA VMP)
Nivel 2 // sintagmas nominales
SN –> TD? AQ* NSN AQ*
Nivel 3 // sintagmas preposicionales
SPR –> SP SN
Nivel 4 // sintagmas verbales
SV –> NSV (SN | SPR)*

2.4 Etiquetado léxico y análisis superficial integrado

En la herramienta también se incorpora un ML que combina diferentes fuentes de conocimiento: probabilidades léxicas, ML de constituyentes sintácticos como SN, SV, etc, y ML de las frases en términos de categorías léxicas y sintácticas. Dicho modelo estadístico, representado como una máquina de estados finitos, es aprendido a partir de corpus anotados con esta información.

El sistema se puede considerar como un traductor a dos niveles. En el nivel superior se representa el ML Contextual para las frases en términos de categorías léxicas y de unidades sintácticas. En el nivel inferior, se modeliza cada unidad sintáctica únicamente en función de las categorías léxicas que la componen. En [Pla00] se describe detalladamente este modelo integrado y se presenta un estudio experimental para la detección de sintagmas nominales sobre el corpus *Penn Treee-Bank*.

Una vez han sido aprendidos estos modelos, y representados como modelos regulares, se construye el ML Integrado realizando una sustitución regular de los niveles inferiores en el superior. Este modelo contempla las posibles concatenaciones de unidades léxicas y sintácticas, además de las probabilidades léxicas aprendidas.

Con este modelo integrado, se puede realizar el proceso de etiquetado léxico y análisis superficial de manera conjunta. El problema consiste en encontrar la mejor secuencia de estados en el modelo para una frase de entrada. Para ello se utiliza el algoritmo de Viterbi, con el que se obteniene el etiquetado léxico y la segmentación en unidades sintácticas de mayor probabilidad.

3 Funcionalidad de la aplicación

A continuación resaltamos las principales tareas que se pueden realizar utilizando el entorno gráfico presentado en este trabajo.

3.1 Edición de etiquetas

La aplicación integra editores de etiquetas léxicas y sintácticas. En principio se han incorporado las etiquetas léxicas definidas en el proyecto *PAROLE* y las utilizadas en el proyecto *Penn Treebank*. Además, estos editores permiten definir cualquier conjunto de etiquetas así como modificar o eliminar las existentes. Las etiquetas léxicas se presentan

Figura 1: Interfaz de representación de una etiqueta léxica

mediante una interfaz donde se reflejan sus rasgos y los distintos valores definidos para cada rasgo (Figura 1).

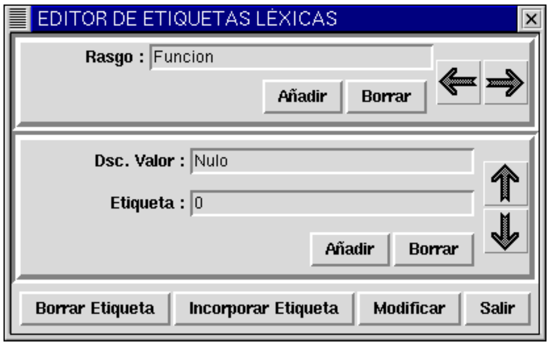


Figura 2: Interfaz del manipulador de etiquetas léxicas

La manipulación de las etiquetas se realiza a través de una segunda interfaz que nos permite añadir, modificar o eliminar rasgos y valores asociados (Figura 2).

El editor de etiquetas sintácticas posee una funcionalidad similar al de las etiquetas léxicas. Estas etiquetas se definen mediante dos parámetros: el nombre y su descripción (Figura 3).

3.2 Edición de gramáticas

Se ha incluido un pequeño editor (Figura 4) que ofrece una serie de facilidades para la construcción y modificación de las gramáticas. En este caso las gramáticas utilizadas

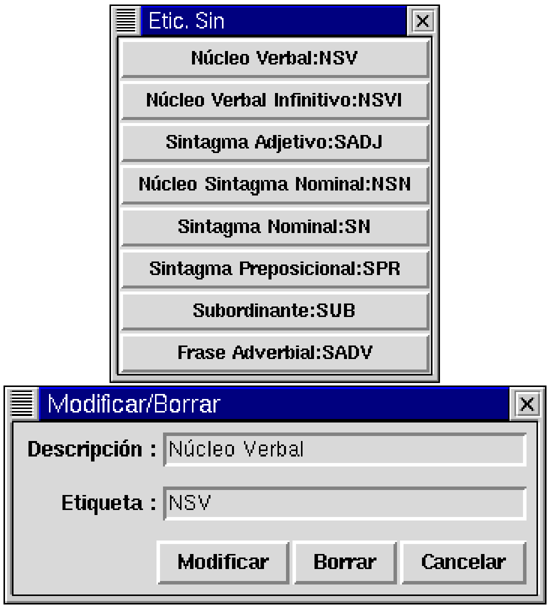


Figura 3: Lista de etiquetas sintácticas e interfaz de manipulación

por el analizador parcial se definen mediante expresiones regulares. Desde la interfaz se pueden escoger los operadores que se han considerado en las definiciones regulares, así como los símbolos del alfabeto que son válidos (sólo se puede usar el conjunto de etiquetas léxicas y sintácticas previamente definidas), evitando la supervisión de las reglas por parte del usuario. Además la comprobación sintáctica se realiza en el proceso de compilación de la gramática.

Figura 4: Editor de gramáticas

3.3 Visualización y corrección del análisis

La herramienta permite llevar a cabo el proceso de análisis de forma secuencial: análisis morfológico, etiquetado léxico y análisis sintáctico, a partir de un corpus de frases no etiquetadas. El usuario puede corregir el análisis después de cualquiera de estas fases, mediante el uso de un *navegador de frases* (Figura 5) que facilita el acceso directo a cualquiera de las oraciones del corpus.

La visualización y edición de oraciones puede realizarse en *modo texto* y en *modo gráfico*. Por ejemplo, la Figura 5 muestra en *modo texto* el resultado del proceso de análisis. Las etiquetas léxicas y sintácticas son sensibles al puntero del ratón y pueden modificarse utilizando las ventanas de edición comentadas anteriormente.

En la Figura 6 puede verse el mismo ejemplo con la presentación en *modo gráfico* del árbol de análisis. De esta forma se facilita tanto la visualización del análisis como la corrección del mismo. Desde el *modo gráfico* es posible añadir, eliminar, modificar o mover cualquier nodo del árbol sintáctico. En la parte derecha de la Figura 6 se muestra un ejemplo que ilustrá cómo se ha completado el análisis sintáctico añadiendo un nuevo nodo (SV) en el árbol. Además de estas características, la herramienta incluye otras para facilitar la visualización del árbol, como la posibilidad de modificar la distancia entre ramas y nodos, la generación de imágenes del árbol sintáctico en formato 'postscript', etc.

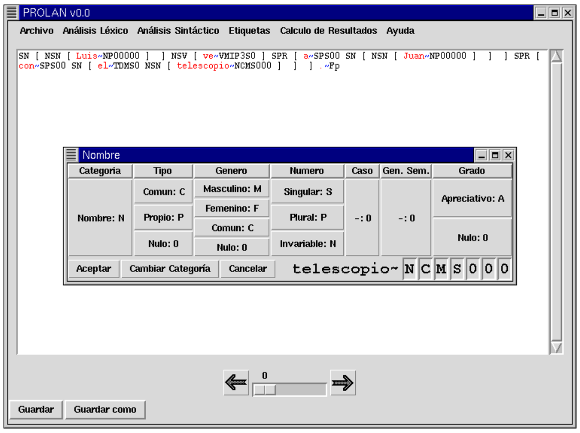


Figura 5: Resultado del análisis con parentizado a izquierdas

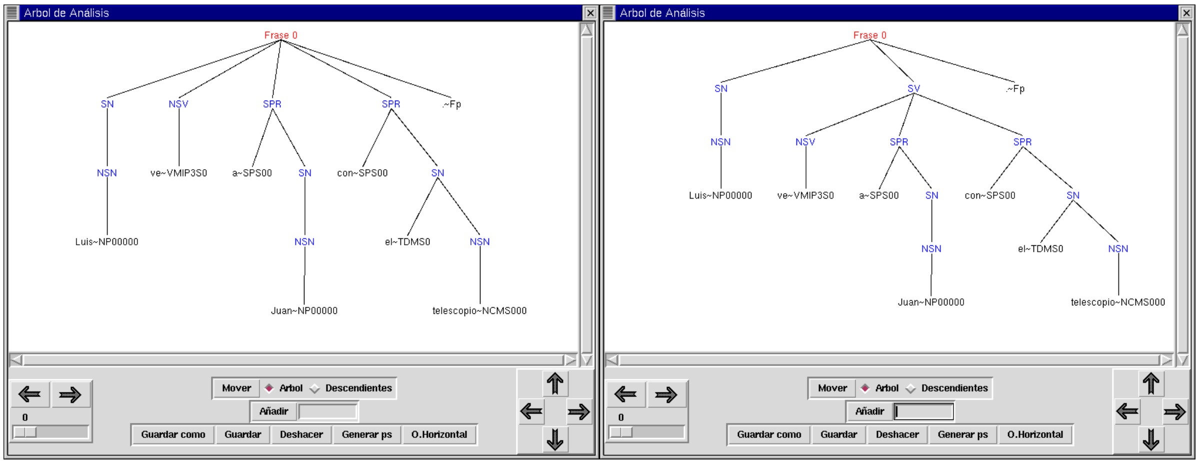
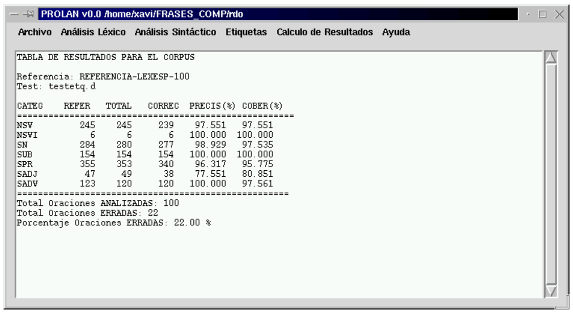


Figura 6: Arbol resultado y árbol corregido

3.4 Evaluación de los analizadores

La herramienta permite evaluar las prestaciones tanto para el etiquetado léxico como para el análisis sintáctico. Para ello se utilizan las medidas habituales: se calcula la precisión de los distintos etiquetadores, así como la precisión y cobertura del análisis sintáctico en lo que respecta a la detección de estructuras no recursivas ('chunks'). En la Figura 7 se muestra un ejemplo de la evaluación de un corpus parentizado.



PROLAN v0.0 /home/xavi/FRASES_COMP/rdo

Archivo Análisis Léxico Análisis Sintáctico Etiquetas Calculo de Resultados Ayuda

TABLA DE RESULTADOS PARA EL CORPUS

Referencia: REFERENCIA-LEXESP-100
Test: testetq.d

CATEG	REFER	TOTAL	CORREC	PRECIS(%)	COBER(%)
NSV	245	245	239	97.551	97.551
NSVI	6	6	6	100.000	100.000
SN	284	280	277	98.929	97.535
SUB	154	154	154	100.000	100.000
SPR	355	353	340	96.317	95.775
SADJ	47	49	38	77.551	80.851
SADV	123	120	120	100.000	97.561

Total Oraciones ANALIZADAS: 100
Total Oraciones ERRADAS: 22
Porcentaje Oraciones ERRADAS: 22.00 %

Figura 7: Resultados de análisis parcial

4 *Conclusiones*

Las principales ventajas de la herramienta presentada son las siguientes:

Facilita la tarea de supervisión de corpora textuales. Estos son útiles como conjunto de entrenamiento en aproximaciones inductivas de modelización del lenguaje. En este sentido, la herramienta puede utilizarse en procesos de 'bootstraping' para construir modelos más fiables y robustos.

Es una herramienta flexible que, aunque inicialmente incorpora dos conjuntos de etiquetas léxicas (*PAROLE* para el castellano y las definidas en *Penn TreeBank* para el inglés), permite que el usuario pueda definirse otros. Además, los nodos del árbol de análisis pueden anotarse con distinta información (roles gramaticales, etiquetas semánticas, etc.), lo cual puede utilizarse para construir corpora de entrenamiento para otras aproximaciones inductivas.

La herramienta permite integrar de manera sencilla cualquier etiquetador léxico o analizador sintáctico que respete los formatos de etiquetado y parentizado sintáctico.

Respecto a posibles mejoras de la herramienta estamos trabajando en los siguientes aspectos:

Incorporación del proceso de entrenamiento para el etiquetado léxico y el análisis integrado.

Incorporación de un analizador morfológico para el inglés, que permita el tratamiento de las palabras desconocidas.

Desarrollo de facilidades para la integración de nuevas herramientas de etiquetado y análisis morfológico y sintáctico.

5 *Agradecimientos*

Este trabajo ha sido parcialmente financiado por el proyecto CICYT (TIC97-0671-C02-01/02).

Agradecemos al *Grup de Processament del Llenguatge* de la Universitat Politècnica de Catalunya por permitirnos el uso del analizador morfológico *MACO*.

Referencias

[Carmona98] Carmona J., Cervell S., Màrquez, L., Martí, M.A., Padró, L., Placer, R., Rodríguez, H., Taulé, M., Turmo, J. "An Environment for Morphosyntactic Processing of Unrestricted Spanish Text". En First International Conference on Language Resource and Evaluation. Granada. Spain. 1998.

[Marti98] Martí, M.A, Rodríguez, H., Serrano, J. "Declaración de categorías morfosintácticas". Documento ITEM nº2. UPC, UB, 1998.

[Marcus93] Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.. "Building a Large Annotated Corpus od English: The Penn Treebank". En Computational Linguistics nº 19, 1993.

[Molina99] Molina, A., Pla, F., Moreno, L., Prieto, N. "APOLN: A Partial Parser of Unrestricted Text". En Papers in Computational Lexicography. Ed. F.Kiefer,G.Kiss y J.Pajzs, (ISBN: 963-9074-20-9), pág. 101-108, 1999.

[Pla98] Pla, F., Prieto, N. " Using Grammatical Inference Methods for Automatic Part-Of-Speech Tagging". En First International Conference on Language Resource and Evaluation. pp.597-601. Granada. Spain. 1998.

[Pla99] Pla, F. and A. Molina. "Etiquetador Léxico basado en Modelos ECGI Extendidos y su aplicación al Corpus BDGEO". En Conferencia de la Asociación Española para la Inteligencia Artificial CAEPIA-TTIA'99. (Murcia, España). Actas del congreso, pp. 36-43 ISBN: 931170-0-5. 1999.

[Pla00] Pla, F., Molina, A. and Prieto, N. "Tagging and Chunking Using Statistical Models". En COLING'2000. Saarbrücken. Germany. 31 July-4 August, 2000.

[Prieto92] Prieto, N., Vidal, E. "Learning Language Models through the ECGI Method". En Speech Communication, N.11, pag. 299-309., 1992.

[Viterbi67] Viterbi, A. "Error bouds for convolutional codes and an asymptotically optimum decodig algorithm". En IEEE Trans. Information Theory, vol. IT-13, pp. 260-269.