Procesamiento del Lenguaje Natural, num. 29 (2002), pp.307-308 recibido 10-07-2002; aceptado 15-07-2002

MPRO - Un programa para el análisis morfológico y sintáctico de textos en español

Johann Haller
IAI –Instituto de Ciencia Aplicada de la Información-
Universidad de Saarland
Martin-Luther-Strasse 14,
66111, Saarbrücken,
Alemania
hans@iai.uni-sb.de

Alexis Donoso
IAI –Instituto de Ciencia Aplicada de la Información-
Universidad de Saarland
Martin-Luther-Strasse 14,
66111, Saarbrücken,
Alemania
alexisdonoso@web.de

Yamile Ramírez
IAI –Instituto de Ciencia Aplicada de la Información-
Universidad de Saarland
Martin-Luther-Strasse 14,
66111, Saarbrücken,
Alemania
yamirasa@web.de

Resumen:

La tarea fundamental del programa MPRO consiste en realizar un análisis morfológico y sintáctico automático de textos españoles técnicos y generales. Las aplicaciones que pueden tener los resultados de estos análisis son de variada índole, pues van desde la traducción automática hasta la indización o elaboración de córpora.

MPRO consiste en una serie de subprogramas, diccionarios, léxicos y gramáticas que interactúan entre sí. Estos componentes pueden ser adaptados a las necesidades del usuario y al tipo de textos con los que se trabaja, sean estos textos especializados o generales.

Tanto el análisis morfológico como el sintáctico son realizados secuencialmente por dos módulos: LESEN y PARSER.

En el primer módulo, LESEN, se lleva a cabo el análisis morfológico. Es decir, es aquí donde las palabras del texto de entrada son analizadas y desambiguadas con informaciones relativas a la morfología, que están contenidas en el **diccionario de morfemas** y en el **fichero de flexiones**. El **diccionario de morfemas** para el español contiene 39.000 entradas. En él, cada palabra está codificada de acuerdo a su estructura y terminación y constituye un alomorfema. En el **fichero de flexiones** se encuentran los valores y características que permiten clasificar cada uno de los alomorfemas de derivación y de flexión que se encuentran en el diccionario de morfemas. Este fichero de datos contiene las reglas necesarias para la formación de palabras. Para el análisis morfológico se utiliza, además, un **diccionario de frecuencia** en el que sólo se introducen aquellas unidades léxicas que no pueden ser analizadas morfológicamente; por ejemplo, las preposiciones, los artículos, los signos de puntuación, los pronombres y algunas locuciones con función preposicional o adverbial.

El texto de entrada del módulo LESEN debe ser un texto en formato ASCII sin marcas SGML o HTML. En primer lugar, el programa intenta identificar cadenas de caracteres separados por espacios y signos de puntuación, con el objeto de reconocer los límites de frase. Para ello, LESEN utiliza las informaciones contenidas en el fichero **limitrules** en donde se establece qué tipo de combinaciones de signos y letras dan inicio a una nueva cadena de caracteres –palabra- o a una nueva frase. Cuando se llega al final de una cadena de caracteres, el programa procede a buscar la forma en el **diccionario de frecuencia**. Si la palabra no se encuentra en este diccionario, entonces LESEN realiza un análisis morfológico para lo cual utiliza los datos presentes en el **diccionario de morfemas** y en el **fichero de flexiones**. En todo caso, antes de segmentar la palabra en morfemas y buscar estos en el **diccionario de morfemas**, LESEN hace uso de las informaciones contenidas en el fichero **wrong,** que contiene alrededor de 80 alomorfemas imposibles de combinar entre sí, ya que su unión da lugar a formas incorrectas, inexistentes en español. De esta forma es posible reducir considerablemente el número de operaciones que debe realizar el programa, así como la duración del análisis morfológico.

El módulo PARSER se encarga de realizar el análisis sintáctico. PARSER trabaja directamente a partir de los resultados de LESEN y consta de una serie de programas y una gramática denominada **newgrammar**. **Newgrammar** es una gramática

ISSN 1135-5948

determinística, procedural y explícita, escrita en un formalismo muy sencillo. La gramática está dividida en cuatro subgramáticas que contienen series de reglas de estructuras frasales. Las reglas no son del todo independientes del contexto, ya que en ellas se puede especificar el contexto inmediato de cada estructura o palabra, lo cual constituye también una forma de restringir las reglas según las necesidades del momento.

La aplicación de las reglas es un proceso sencillo. El programa utiliza las subgramáticas en secuencia para analizar la frase u oración. Cuando todas las reglas de la primera subgramática han sido probadas y al menos una funciona, el programa aplica una vez más la misma subgramática para asegurarse de que sólo esa regla es la correcta. En caso de que no funcione ninguna regla de la primera subgramática, el programa pasa a la siguiente.

Las reglas se aplican en el orden en que se encuentran en las subgramáticas. Cuando se cumplen todas las condiciones de una regla, el programa construye -a partir de la entrada original de la oración- la estructura sintáctica descrita en la regla. Dado que las ambigüedades sintácticas no se resuelven en esta fase –como podría ocurrir al realizarse un *backtracking*–, es posible que algunos análisis resulten falsos. No obstante, con una secuencia adecuada de las reglas y sus respectivos contextos se pueden evitar análisis erróneos.

Gracias a la sencillez del proceso, el análisis de los textos es sumamente rápido y, generalmente, se obtienen suficientes grupos de palabras que más tarde pueden utilizarse en la indización automática o en la traducción informativa. Cabe señalar que el módulo PARSER se encuentra aún en desarrollo y que, en su estado actual, no está en condiciones de ofrecerle al usuario un análisis sintáctico completo de cualquier oración castellana.

Demostración:

Para la demostración de MPRO se ha previsto realizar un análisis morfológico y sintáctico de textos periodísticos tomados del periódico español *El País*. Los resultados obtenidos serán mostrados y analizados, al tiempo en que se describirán los diferentes componentes del programa. Se estima que para la presentación del producto se necesitarán 30 minutos.