Ingeniería del Lenguaje Natural

Práctica 8. Análisis parcial

1. Introducción

Muchas aplicaciones de PLN no necesitan desarrollar análisis sintácticos completos de los textos para ser procesados. En la mayoría de las ocasiones, con un análisis superficial es suficiente para la aplicación que se está desarrollando.

Entre estas aplicaciones, una de las que más uso hace del análisis sintáctico parcial es la Extracción de Información. Un sistema de extracción de información que, por ejemplo, quiera extraer nombres de personas y nombres de organizaciones (un sistema de extracción de entidades nombradas –*Named Entities*-) le basta con detectar en el texto los sintagmas nominales, sin tener que desarrollar el análisis sintáctico completo del texto.

En esta práctica se va a desarrollar un sencillo analizador sintáctico parcial (*chunker*). Para ello se va a utilizar NLTK. Éste permite desarrollar un *chunker* a partir de expresiones regulares.

2. Desarrollo

1. Leer los apartados 7.1. al 7.4 de Bird et al *Introduction to Natural Language Processing* (disponible en la web del NLTK)

2. Analiza en el corpus conll2000 *train* qué tipos de sintagmas nominales aparecen.

3. Desarrollar expresiones regulares que analicen únicamente sintagmas nominales a partir de la etiqueta de PoS. Pruébalo con el corpus conll2000 *train* (formato no analizado).

4. Al final, analiza el corpus conll2000 *test*.

5. Evalúa la precisión de la gramática según se indica en el apartado 7.4.2.

3. Entrega

1. La gramática (el conjunto de expresiones regulares)

2. El corpus conll2000 test analizado

3. El valor de la evaluación (punto 5)