Joan Curto Alonso 183670
Juan García del Muro Navarro 164880
Lluis Hysa 160317

Práctica 1 de Procesamiento de Lenguaje Natural

En esta primera práctica nos dedicaremos a etiquetar un texto utilizando el método más sencillo, el de asignarle a cada token el tag con el que dicho token que aparezca más en el training set, conocido por nuestro programa antes de comenzar.

Estructura

Ya que esta práctica requiere 3 pasos, hemos dividido el código en tres funciones que hacen referencia a cada uno de dichos pasos, a las que hemos llamado primerPas(), segonPas() y tercerPas(). Además también hemos definido varias funciones con la finalidad de encapsular tareas específicas y de esta forma simplificar el código lo más posible.

En el primer paso leemos el archivo corpus.txt que se trata de nuestro training set. Dicho archivo incluye una lista de tokens etiquedados con un tag determinado. Con dicha información generamos el archivo lexic.txt, en el que guardamos una lista de tokens, cada uno con su tag correspondiente y el número de veces que aparece en el corpus con ese tag.

En el segundo paso tomamos los archivos test_1.txt y test_2.txt y proseguimos a etiquetarlos. Lo hacemos basándonos en la información recopilada en el primer paso, asignando a cada token el tag con el que más apareciera dicho token en el archivo lexic.txt. Los resultados los guardamos en los archivos tagged_test_1.txt y tagged_test_2.txt.

En el tercer paso, simplemente comparamos los resultados del segundo paso con los archivos gold_standard_1.txt y gold_standard_2.txt, que contienen cada uno de los tokens etiquetados de forma correcta. El resultado de

dicha comparación es que en la etiquetación del archivo test_1.txt nuestro programa consigue un porcentaje de acierto del 88% y en la etiquetación del archivo test_2.txt consigue un 81%.

Siguientes pasos

A pesar de que 88% y 81% son resultados respetables para un algoritmo de etiquetación tan simple como este, tiene un problema de base. La lógica puede llevar a pensar que para mejorar el algoritmo simplemente deberíamos ampliar el training set del que se alimenta el etiquetador. Sin embargo, dicha medida no tendría un impacto significativo en el porcentaje de acierto. La mejor manera de alcanzar porcentajes cercanos al 100% sería implementar un algoritmo de procesamiento de lenguaje más complejo que tuviera en cuenta el contexto en el que cada token ha sido etiquetado con un tag determinado y extraer información de este, como su significado.