PRÁCTICA 1: Detección de Entidades Nombradas

El objetivo de esta práctica es utilizar un etiquetador de entidades nombradas, en español, y evaluar los resultados obtenidos. Se puede utilizar el etiquetador de entidades nombradas de Spacy, que cuenta con un modelo preentrenado para el español. Es decir, no se pide entrenar el modelo, solo utilizar las herramientas. El objetivo es ver la utilidad de las herramientas para procesamiento del lenguaje aplicadas a un problema planteado en una competición científica, y analizar los resultados, estudiando las causas de los casos de error y cómo pueden mejorarse. Para realizar esta práctica es necesario haber estudiado previamente los contenidos de los temas 1 y 2.

Utilizaremos textos de prueba pertenecientes al conjunto de test del corpus conll2002 en español:

Fichero esp.testb (https://www.clips.uantwerpen.be/conll2002/ner/data/).

De esta forma contaremos con unas anotaciones de referencia con las que podemos comparar las del etiquetador utilizado en la práctica y evaluar su calidad. La tarea de evaluación para la que se propuso este corpus se describe en el siguiente artículo:

> *Erik F. Tjong Kim Sang: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Taipei, Taiwan, 2002*
> https://www.aclweb.org/anthology/W02-2024

El formato de los datos del corpus de referencia utiliza una línea para cada palabra, seguida de la anotación de la entidad en formato IOB. Los tipos de entidades considerados son:

- PER: Persona
- ORG: Organización
- LOC: Localización
- MISC: Miscelánea

Por ejemplo:

> *La B-LOC*
> *Coruña I-LOC*
> *, O*
> *23 O*
> *mayo O*
> *(O*
> *EFECOM B-ORG*
> *) O*
> *. O*

Es necesario extraer el texto plano a partir de los ficheros para aplicarle el etiquetador de Spacy.

Finalmente, para la evaluación de los resultados se debe calcular la precisión, cobertura y medida-F. Podemos utilizar el script proporcionado para la tarea de conll2002 y disponible en el entorno virtual: *conlleval.py*

El formato de los datos que espera este programa es el siguiente (palabra, etiqueta de referencia, etiqueta asignada): *word gold pred*

Por ejemplo:

La B-LOC B-LOC
Coruña I-LOC I-LOC
, O O
23 O O
mayo O O
(O O
EFECOM B-ORG B-ORG
) O O
. O O

Documentación a entregar:

Subir un archivo comprimido que contenga lo siguiente:

- Documento PDF (en ningún caso se entregarán archivos tipo notebook o el pdf generado a partir de un notebook) con la siguiente información:
    - Descripción de la tarea conll2002 (comentar los principales puntos y resultados del artículo) y los datos de evaluación.
    - Descripción del código desarrollado, herramientas utilizadas, etc.
    - Textos de prueba utilizados.
    - Resultados de evaluación del etiquetado.
    - Análisis de los errores de etiquetado y sus causas. Se valorará la introducción de mejoras/modificaciones en el etiquetado que mejoren los resultados.
- Directorio que contenga el código fuente y los ficheros generados.

OPCIONAL

Anotar manualmente algunos textos que contengan abundantes entidades nombradas (por ejemplo, documentos de economía, o textos médicos, etc.), y con suficiente longitud para que los resultados sean significativos. Para la anotación se puede seguir el formato que se desee, siempre que sea fácil de leer. Después se debe trasformar al formato IOB, con la misma forma que tiene la parte obligatoria. Se debe evaluar la anotación automática de los textos, comparándola con la

anotación manual, y documentar y comentar (documento pdf, en ningún caso se entregarán archivos notebook ni el pdf generado a partir de un notebook):

- El formato de anotación utilizado.
- Los textos utilizados sin y con anotaciones.
- Los resultados comparativos de la anotación automática y manual.
- Comparativa con los resultados de la parta obligatoria.
- Reflexiones sobre los resultados y comparativas.

También puede estudiarse la anotación de textos en inglés.

A título orientativo, la valoración de la parte obligatoria será de entre 5 y 9 puntos, siendo imprescindible obtener al menos 5 para aprobar. La valoración de la parte opcional será de entre 0 y 2 puntos en función del grado de aportación. La suma de la nota de ambas partes nunca será superior a 10. Al haber bastante libertad para realizar análisis y mejoras de los resultados del etiquetado, la evaluación de cada parte va a depender del grado de aportación.