

# EVALUACIÓN DE ETIQUETADORES MORFOSINTÁCTICOS PARA EL ESPAÑOL

## 1. Objetivos del trabajo:

- Conocer distintas herramientas de etiquetado morfosintáctico del paquete NLTK
- Aplicar la metodología de validación cruzada para entrenar y evaluar etiquetadores morfosintácticos.
- Construir modelos de etiquetado morfosintáctico para el español a partir de corpus anotados
- Comparar las prestaciones de distintos etiquetadores incluyendo intervalos de confianza

## 2. Descripción del trabajo:

El trabajo consiste en la evaluación de las prestaciones de distintos etiquetadores morfosintácticos para un corpus del español (*cess-esp*). En esta experimentación se estudiará cómo afectan diversos parámetros a las prestaciones del sistema: el tamaño del corpus de entrenamiento, el método de suavizado para las palabras desconocidas, el juego de categorías morfosintácticas utilizado, etc.

Además, se compararán las prestaciones de distintos etiquetadores morfosintácticos basados en distintos paradigmas de aprendizaje.

Para ello, se utilizará el paquete NLTK que implementa diferentes etiquetadores morfosintácticos.

La evaluación de los etiquetadores se realizará mediante una **validación cruzada** sobre 10 particiones del corpus. Esta metodología consiste en dividir el corpus en 10 partes de similar tamaño, y ejecutar diez experimentos. En cada ejecución se toman 9 partes como entrenamiento y 1 como prueba, de manera que la parte de prueba siempre sea diferente.

La evaluación de las prestaciones del etiquetador es el resultado de calcular la media de la precisión de etiquetado (*accuracy*) sobre las distintas particiones.

$accuracy = \text{número palabras etiquetadas correctamente del test} / \text{número total de palabras del test}$

IMPORTANTE: Los resultados de *accuracy* obtenidos deberán presentarse siempre con el correspondiente intervalo de confianza.

**Tarea 1: Evaluación del etiquetador 'hmm' sobre el corpus 'cess-esp' utilizando el juego de categorías completo y reducido.**

Utilizando el etiquetador **hmm** basado en modelos de Markov, se realizará una validación cruzada sobre 10 particiones del corpus. Barajar el corpus antes de realizar las particiones. Presentar los resultados en forma de tabla y gráficamente, incluyendo los intervalos de confianza.

## **Tarea 2: Evaluación de las prestaciones del etiquetador respecto a la cantidad de datos de aprendizaje.**

Se trata de estudiar cómo varían las prestaciones del etiquetador *hmm* cuando varía el tamaño del corpus de aprendizaje. Para este experimento se dividirá el corpus de entrenamiento en 10 partes de tamaño similar. La partición 10 se tomará como test, y las 9 particiones restantes se tomarán como entrenamiento. En cada ejecución, se irá incrementando sucesivamente el tamaño del corpus de entrenamiento, manteniendo fija la partición de test.

**Importante:** Para esta tarea **no es necesario** realizar la validación cruzada.

## **Tarea 3: Evaluación del método de suavizado para palabras desconocidas para el etiquetador *tnt*.**

El etiquetador *tnt* por defecto no incorpora un método de suavizado para las palabras desconocidas. Utiliza un método basado en los sufijos de las palabras para construir un modelo para las palabras desconocidas (**Affix Tagger**). En base al sufijo de la palabra desconocida le asigna una categoría morfosintáctica. Este método funciona razonablemente bien para el inglés.

En concreto, se trata de estudiar diferentes longitudes del sufijo (número de letras que se tienen en cuenta) y estudiar cómo varían las prestaciones del etiquetador. Una vez se haya decidido el sufijo que mejores prestaciones proporciona, incorporarlo como modelo de suavizado al etiquetador *tnt* y comprobar si aumenta sus prestaciones.

## **Tarea 4: Evaluación del resto de etiquetadores.**

Se deberán utilizar otros paradigmas de etiquetado. Como mínimo el etiquetador de Brill y algún otro como, CRF, perceptron. Se deberá realizar una comparativa de prestaciones respecto a los etiquetadores *tnt* y *hmm*, utilizando el juego de categorías reducido.

Cuando se utilice el etiquetador de Brill, probar con diferentes etiquetados iniciales, por ejemplo probar con Unigram Tagger y con *hmm* tagger.

La comparación puede ser sólo de una partición, si el coste temporal de la validación cruzada requiere mucho tiempo.

## **Tarea 5: Opcional (10% del trabajo)**

Se deberá realizar al menos el estudio de una de las siguientes herramientas. En dicho estudio se debe contemplar aspectos como: facilidad/problemas de instalación, facilidad de uso, documentación, funcionalidad, etc. Con cada herramienta estudiada se deberá realizar el etiquetado morfosintáctico del texto del fichero *Alicia.txt*. Se debe entregar un fichero tipo texto con el formato: *palabra/etiqueta*. Además, se debe explorar la posibilidad de etiquetar el fichero *Alicia.txt* con entidades nombradas (NE).

**5.1 Evaluación de la herramienta Spacy.** Esta herramienta de libre distribución se puede obtener en la siguiente dirección <https://spacy.io/>

**5.2 Evaluación de la herramienta Stanza.** Esta herramienta de libre distribución se puede obtener en la siguiente dirección <https://stanfordnlp.github.io/stanza/>

## **Tarea 6: Opcional (10% del trabajo)**

Entrenar el *Spacy* con el corpus *cess-esp* para realizar *POS tagging* en español y obtener las prestaciones del etiquetador en términos de *accuracy*.

### **Características del trabajo (se recomienda hacerlo en grupo de dos personas).**

- Se debe presentar una memoria que describa el trabajo realizado: tanto en la fase de ajuste (si se ha realizado) como los resultados obtenidos sobre el conjunto/s de test (presentados en forma de tablas y gráficas), así como el análisis y comparación de los resultados obtenidos.
- El código Python desarrollado.
- El trabajo se debe presentar públicamente. Preparad una breve presentación del trabajo realizado, resaltando los aspectos que se consideren más importantes, o las dificultades encontradas. En la presentación deberán intervenir los dos miembros del equipo.