UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MIARFID
Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

DSIC
DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital
Universitat Politècnica de València

Evaluación de distintos etiquetadores morfosintácticos para el español

Lingüística Computacional

Autor: Jose Gómez Gadea
Juan Antonio López Ramírez

Curso 2019-2020

Introducción

El objetivo de esta memoria 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.

Tarea 1

En esta sección se pretende, utilizando el etiquetador hmm basado en modelos de Markov, realizar una validación cruzada sobre 10 particiones del corpus. Se ha barajado el corpus antes de realizar las particiones.

Se pueden observar los resultados de esta tarea en la tabla 1 y en la figura 1, donde también se incluyen los intervalos de confianza.

Partición para test	**Tasa de acierto (%)**	**Intervalo de confianza (%)**
1	92.31	±2,12
2	92.80	±2,06
3	92.38	±2,11
4	92.71	±2,07
5	92.16	±2,14
6	92.54	±2,09
7	92.68	±2,07
8	92.84	±2,05
9	92.31	±2,12
10	92.96	±2,04

Tabla 1: Tabla con los resultados de aplicar hmm sobre el corpus barajado.

En la tabla se puede observar como se mantiene el ratio de acierto bastante similar en todas las particiones. Esto se debe a que las muestras tanto de entrenamiento como de test se han cogido del mismo corpus y que, al estar barajado, nos

aseguramos de que las distintas particiones realizadas sobre él van a ser bastante similares en cuanto a variedad de contenido.

Tarea 2

En esta tarea se pretende estudiar cómo varían las prestaciones del etiquetador hmm cuando varía el tamaño del corpus de aprendizaje. Para este experimento se ha dividido el corpus de entrenamiento en 10 partes de tamaño similar. La partición 10 se ha tomado como test, y las 9 particiones restantes se han tomado como entrenamiento. En cada ejecución, se ha incrementado sucesivamente el tamaño del corpus de entrenamiento, manteniendo fija la partición de test.

Se pueden observar los resultados de esta tarea en la tabla 2 y en la figura 2, donde también se incluyen los intervalos de confianza.

Partición añadida entrenamiento	**Tasa acierto (%)**	**Intervalo confianza (%)**
1	83.79	±2,94
2	87.06	±2,67
3	88.84	±2,51
4	89.86	±2,40
5	90.59	±2,33
6	91.11	±2,27
7	91.76	±2,19
8	92.20	±2,14
9	92.52	±2,09

Tabla 2: Tabla aplicando HMM con el corpus barajado incrementando la cantidad de elementos de entrenamiento.

Se puede observar como la tasa de acierto se ha ido incrementando a medida que se han ido añadiendo particiones al entrenamiento. Esto se debe a que el etiquetador ha podido generar un modelo más completo, que tiene en cuenta un mayor número de posibilidades.

Tarea 3

Dado que el etiquetador tnt por defecto no incorpora un método de suavizado para las palabras desconocidas, se pretende utilizar 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, se le asigna una categoría morfosintáctica. Este método funciona razonablemente bien para el inglés.

En nuestro caso, se va a estudiar cómo varían las prestaciones del etiquetador según la utilización de diferentes longitudes del sufijo (número de letras que se tienen en cuenta). Comentar que para las diferentes pruebas, se va a utilizar también validación cruzada sobre 10 particiones del corpus.

Se pueden observar los resultados de esta tarea en la tabla 3 y en la figura 3, donde también se incluyen los intervalos de confianza.

Número de sufijo(s) utilizados	Tasa de acierto (%)	Duración (segundos)
0	90.17	96.29
1	92.95	96.60
2	93.88	95.30
3	94.73	95.54
4	94.45	92.95

Tabla 3: Tabla aplicando TNT con el corpus barajado, comparando la media de aciertos por cada tamaño de sufijo y el tiempo que ha tardado en generar el modelo y clasificar la muestra.

Se puede observar que el etiquetador ha ido elevando su tasa de acierto a medida que se ha ido cogiendo más caracteres para el suavizado, hasta llegar al número de caracteres 3. A partir de aquí, el etiquetador ha sacado incluso una puntuación menor. Esto se debe a que a partir del cuarto caracter, la clasificación de los sufijos empeora.

Tarea 4

Se pretende utilizar otros paradigmas de etiquetado para evaluar su precisión. Se ha utilizado el etiquetador de Brill y perceptron. Se ha realizado una comparativa de prestaciones respecto a los etiquetadores tnt y hmm, utilizando el juego de categorías reducido.

Para el etiquetador de Brill, se ha probado con diferentes etiquetados iniciales, en este caso Unigram Tagger y hmm tagger. Para la comparación se ha realizado validación cruzada.

Etiquetador utilizado	Tasa de acierto (%)	Duración (segundos)
Brill con Unigram Tagger	89.63	138.82
Brill con HMM Tagger	92.70	1292.45
Perceptron	96.75	1163.31
TNT	90.21	1217.10
HMM	92.63	283.75

Tabla 4: Tabla aplicando diferentes etiquetadores con el corpus barajado, comparando la media de aciertos de cada uno y el tiempo que ha tardado en generar el modelo y clasificar la muestra.

Tarea 6

Esta tarea se ha realizado en Linux por la facilidad de instalación de Freeling en dicho sistema operativo. Se ha intentado utilizar previamente en Windows, pero ha dado muchos problemas por lo que se ha descartado. Para la preparación en Linux se ha instalado el .deb que se ofrece en su propia Web (aunque también se puede instalar mediante pip).

Nos ha parecido que la librería tiene un funcionamiento incluso más sencillo que NLTK, la problemática ha llegado al intentar obtener información de la respuesta de Freeling en formato XML. Dado que los ejemplos y gran parte de la documentación de la web ha sido eliminada de sus servidores (o están caídos), no ha sido posible conocer una descripción sobre los métodos que aporta la librería. Esta problemática nos ha llevado a guiarnos de foros independientes y documentaciones a medias recogidas de diversas páginas web.

Finalmente, se ha tenido que inspeccionar la respuesta en XML mediante un *parser* independiente a Freeling, aunque se ha podido obtener todas las palabras con su etiquetado morfosintáctico sin mayores complicaciones.

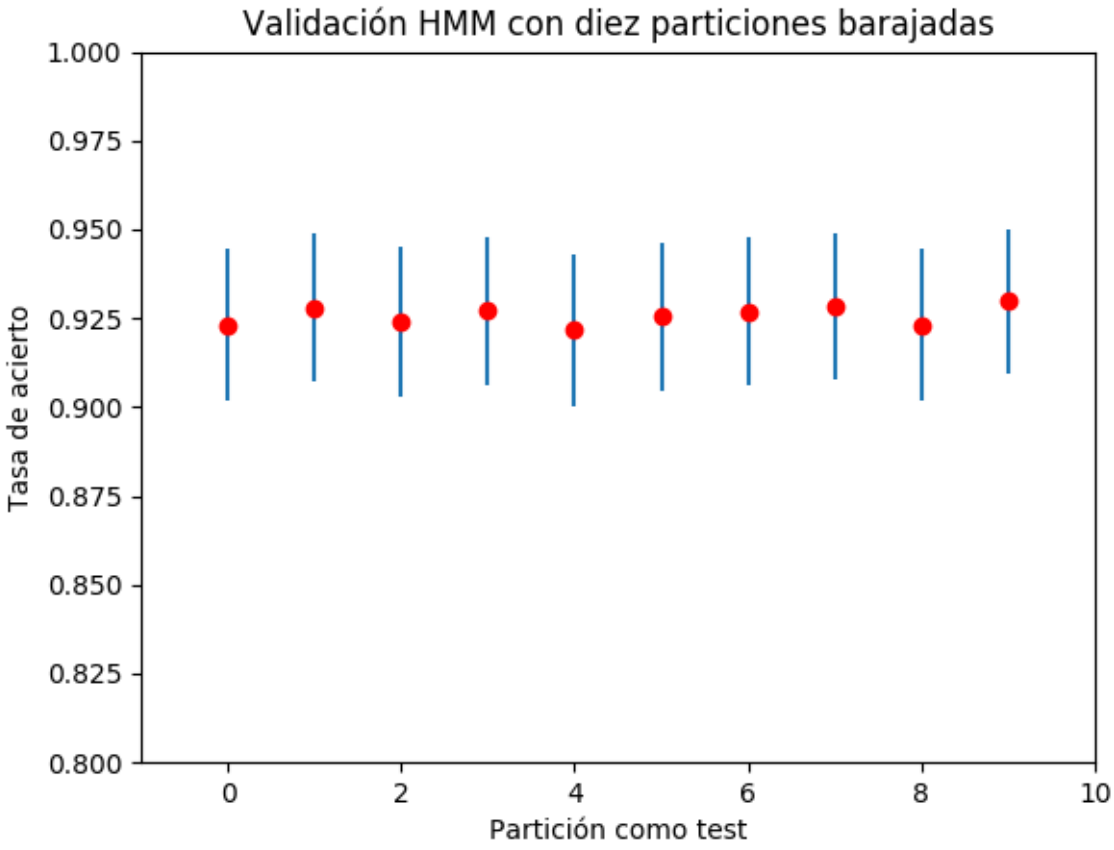

Figura 1: Gráfica con los resultados de aplicar HMM sobre el corpus barajado.

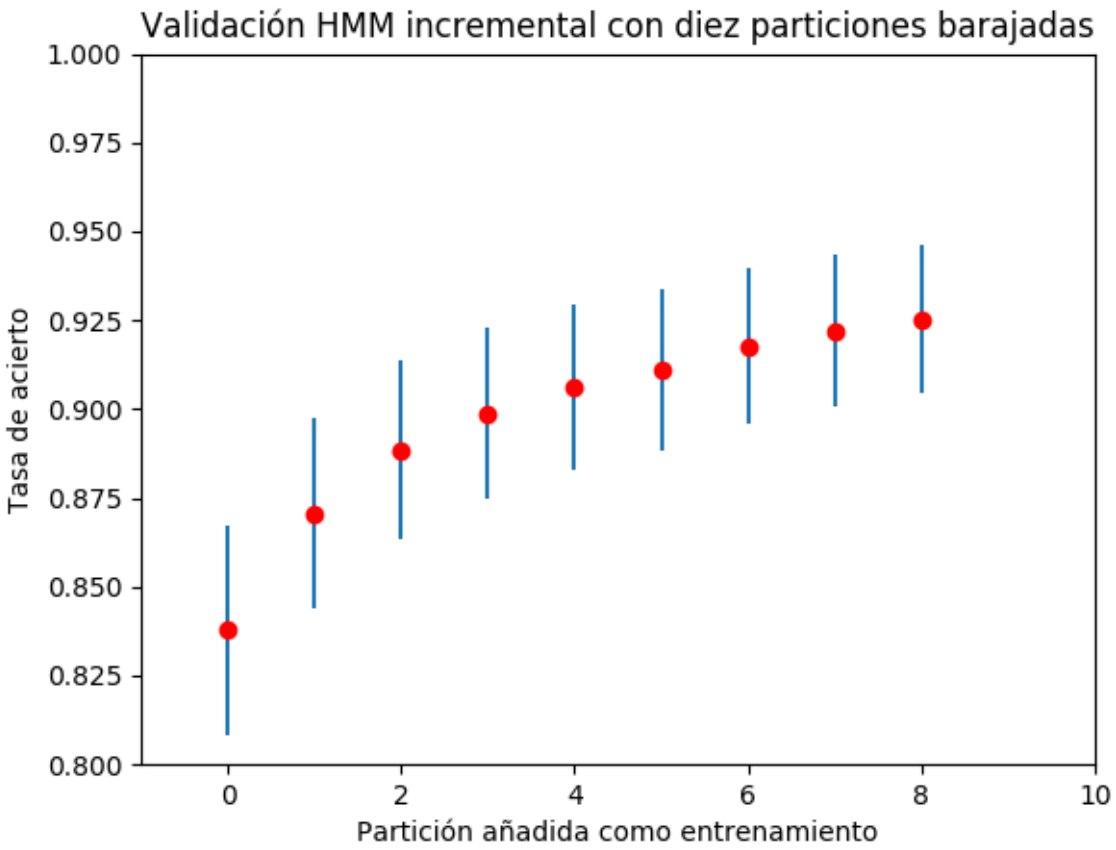

Figura 2: Gráfica con los resultados de aplicar HMM barajando el corpus y variando la cantidad de elementos de entrenamiento.

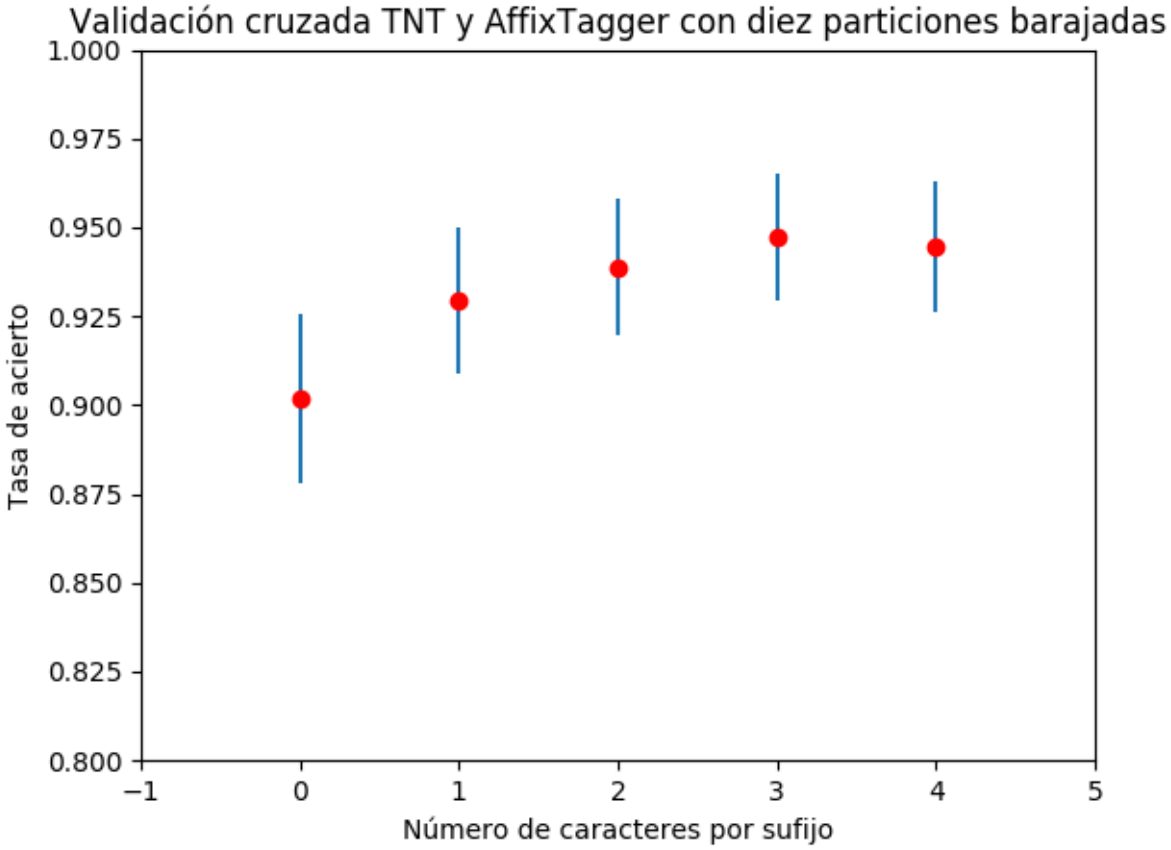

Figura 3: Gráfica con los resultados de aplicar TNT barajando el corpus y variando la cantidad de sufijos con Afflix Tagger.

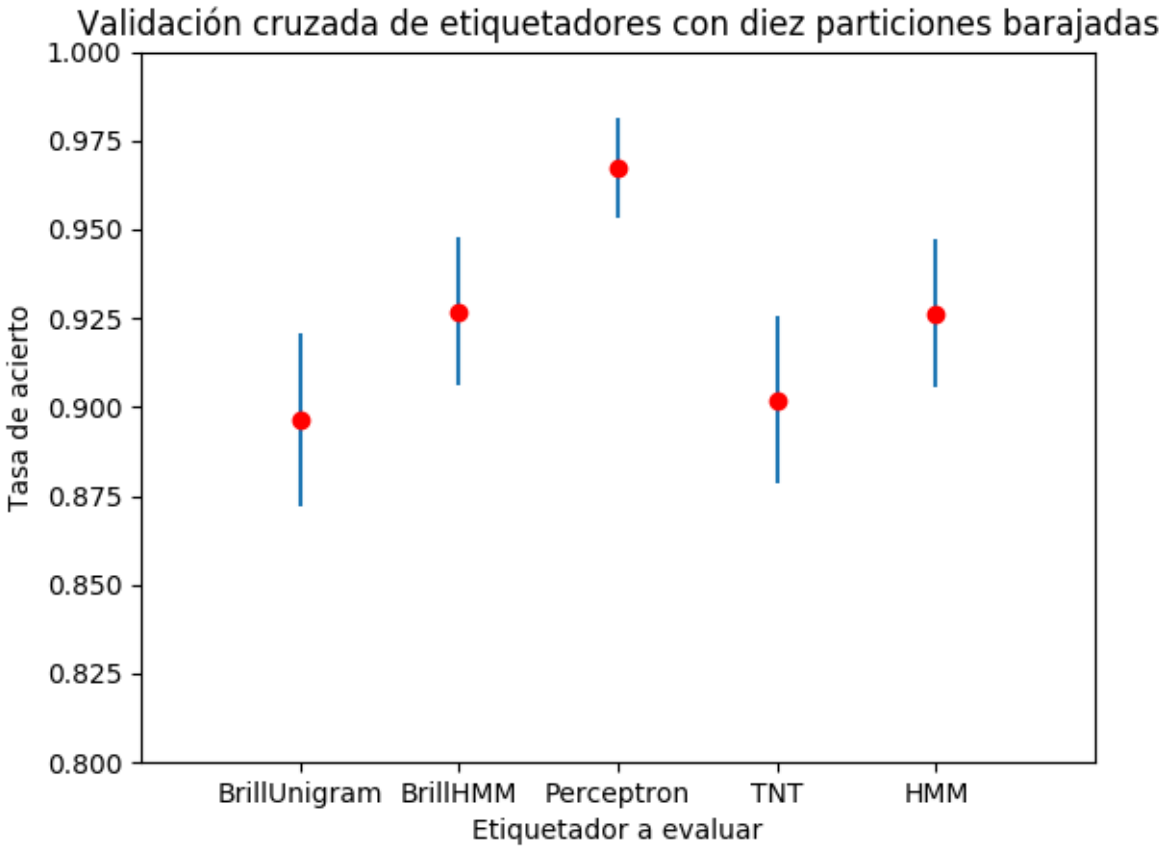

Figura 4: Gráfica con los resultados de aplicar los distintos etiquetadores con el corpus barajado.