Evaluación de etiquetadores morfosintácticos para el español

Jose Arias Moncho
Victoria Beltrán Domínguez

ÍNDICE

1. Introducción

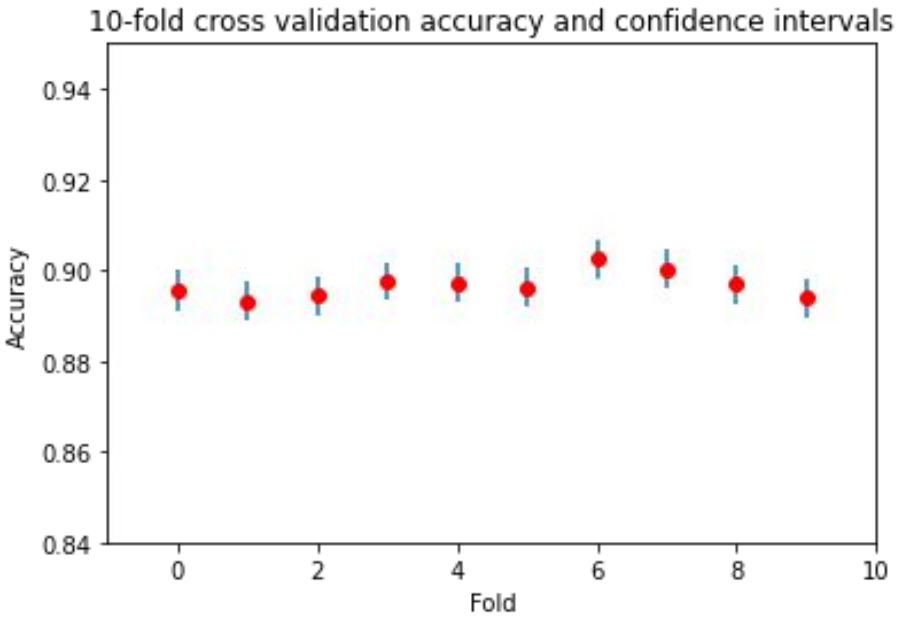
- Corpus cess-esp
  - 6030 frases etiquetadas
  - [('El', 'da0ms0'), ('grupo', 'ncms000'), ('estatal', 'aq0cs0'), ('Electricité_de_France', 'np00000'), ...]

- Barajado

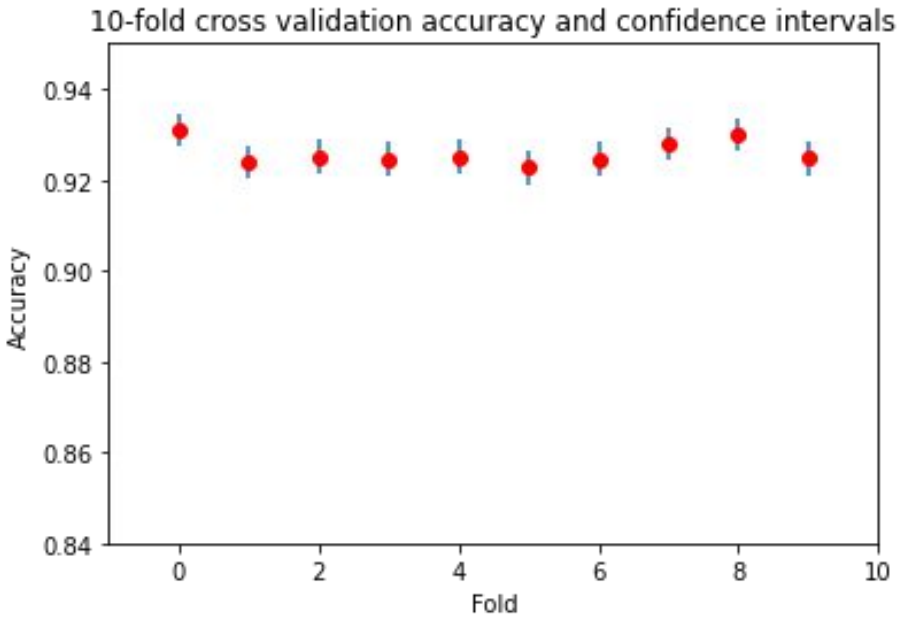
- Validación cruzada con 10 particiones

1. Evaluación del etiquetador hmm

Juego de categorías completo

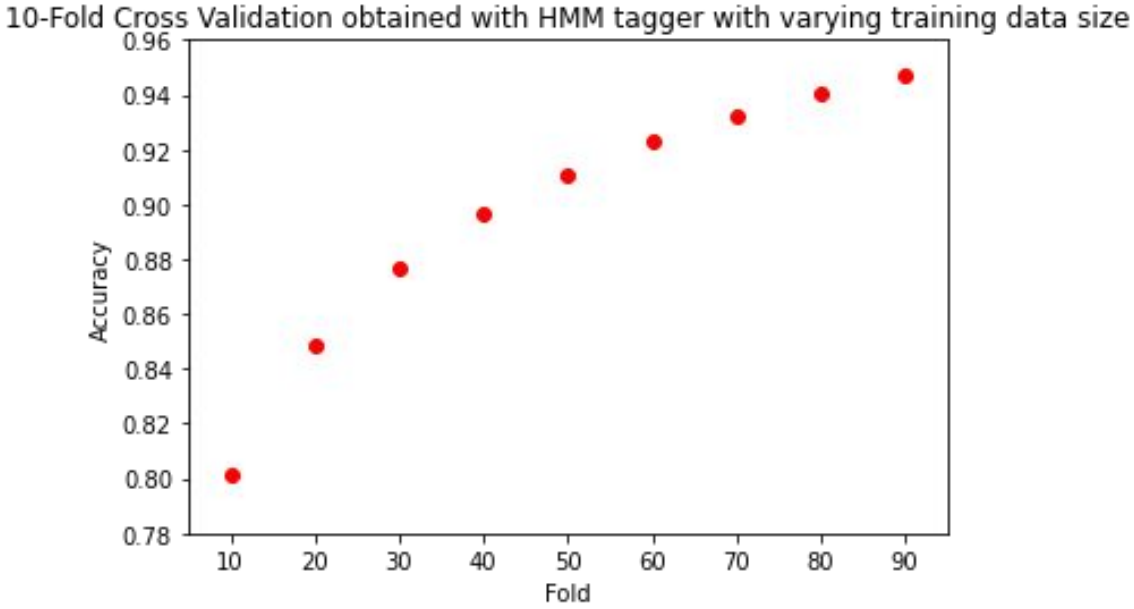


Juego de categorías reducido



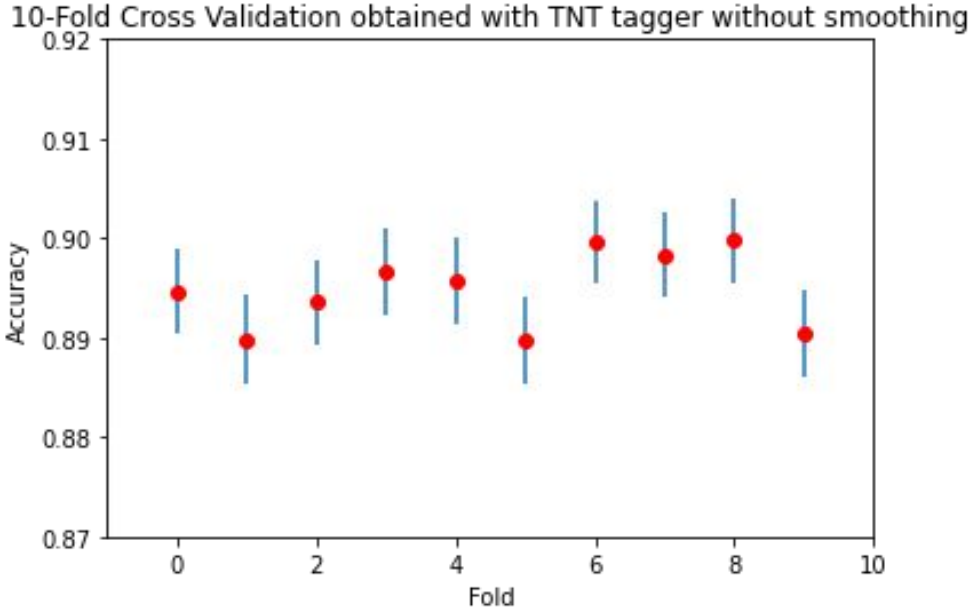
2. Evaluación de las prestaciones del etiquetador respecto a la cantidad de datos de aprendizaje

- Etiquetador hmm
- División del corpus en 10 partes



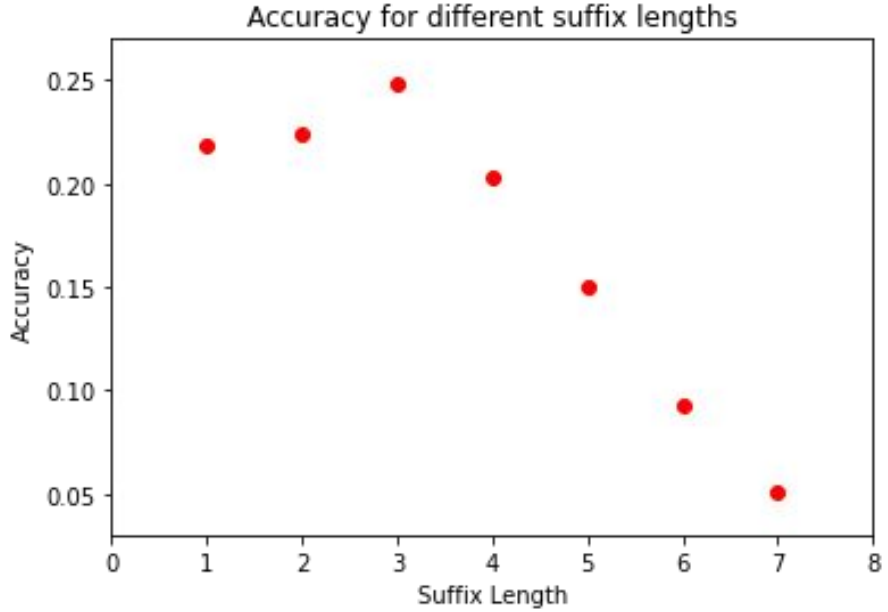
3. Evaluación del método de suavizado para palabras desconocidas (I)

- Etiquetador tnt
- Validación cruzada 10 particiones



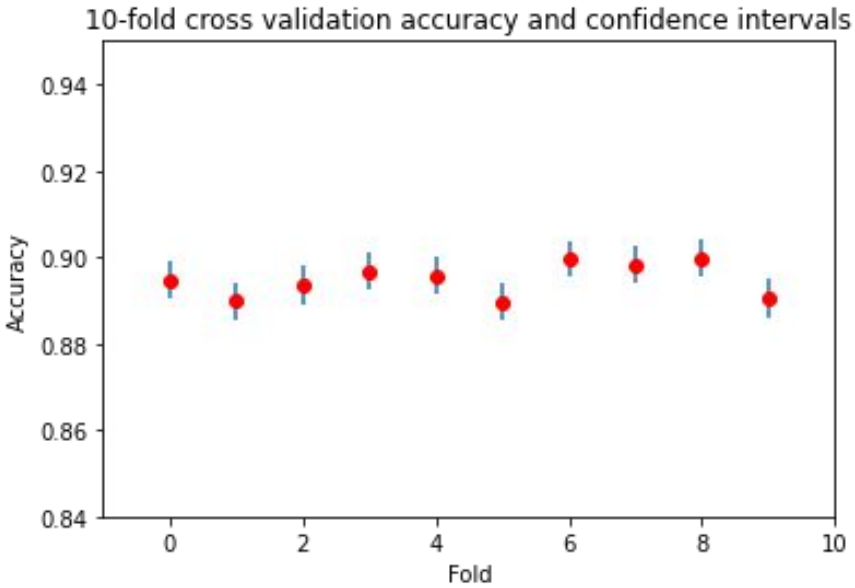
3. Evaluación del método de suavizado para palabras desconocidas (II)

- Affix Tagger
- Tamaño de sufijo (|suf|={ 1, 2, 3, 4, 5, 6, 7 }

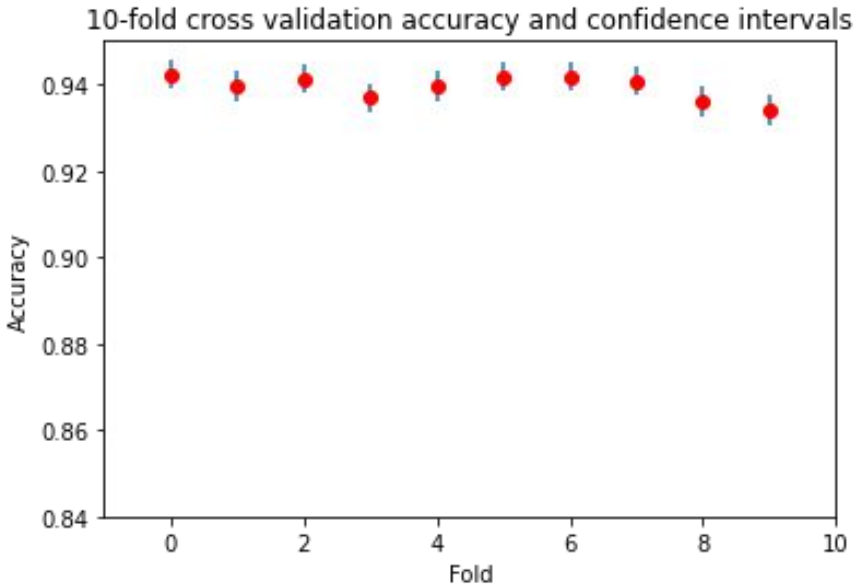


3. Evaluación del método de suavizado para palabras desconocidas (III)

Sin suavizado

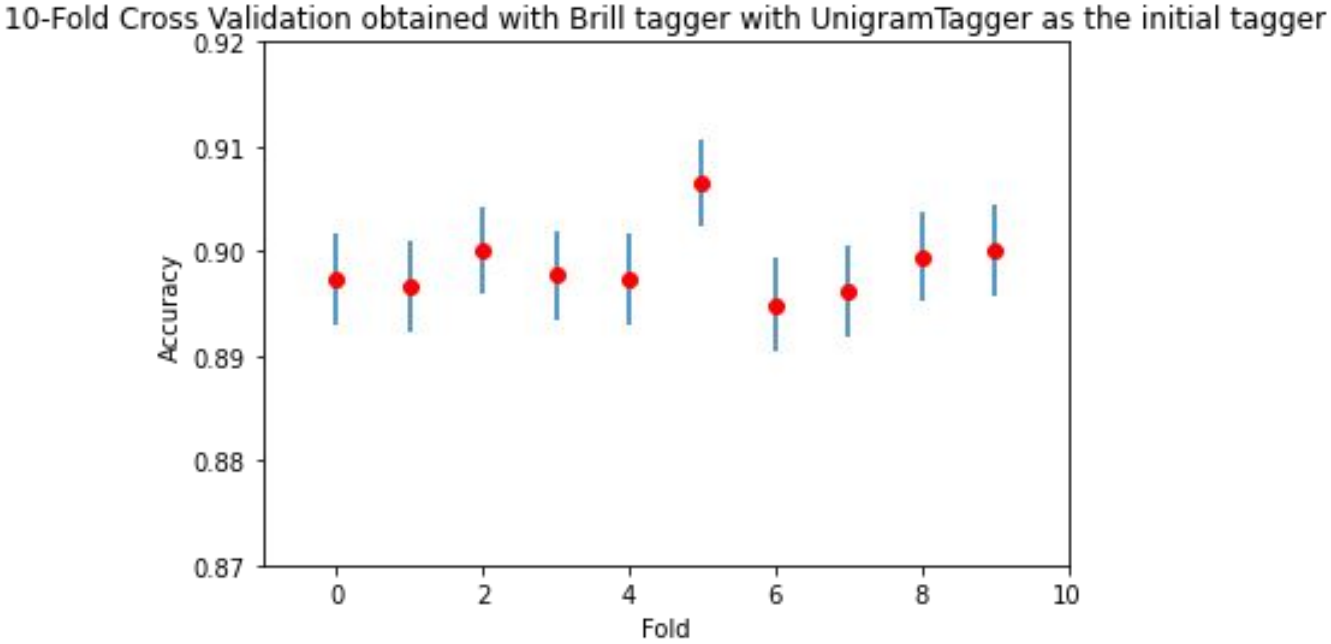


Con suavizado



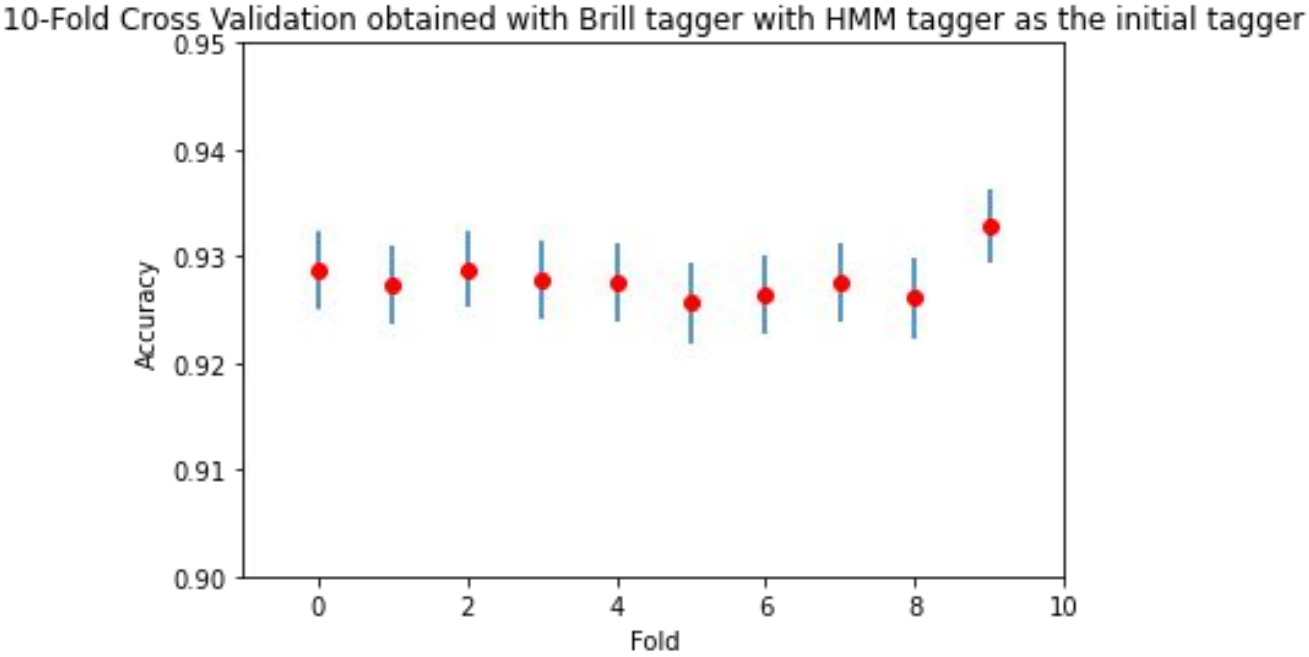
4. Evaluación del resto de etiquetadores (I)

Brill Tagger + Unigramas



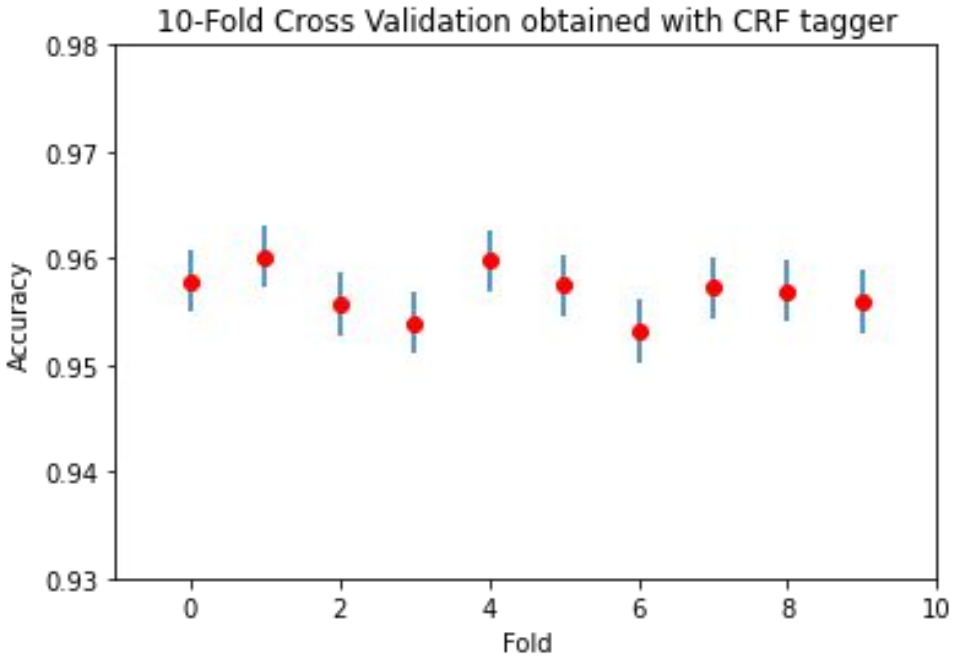
4. Evaluación del resto de etiquetadores (II)

Brill Tagger + HMM



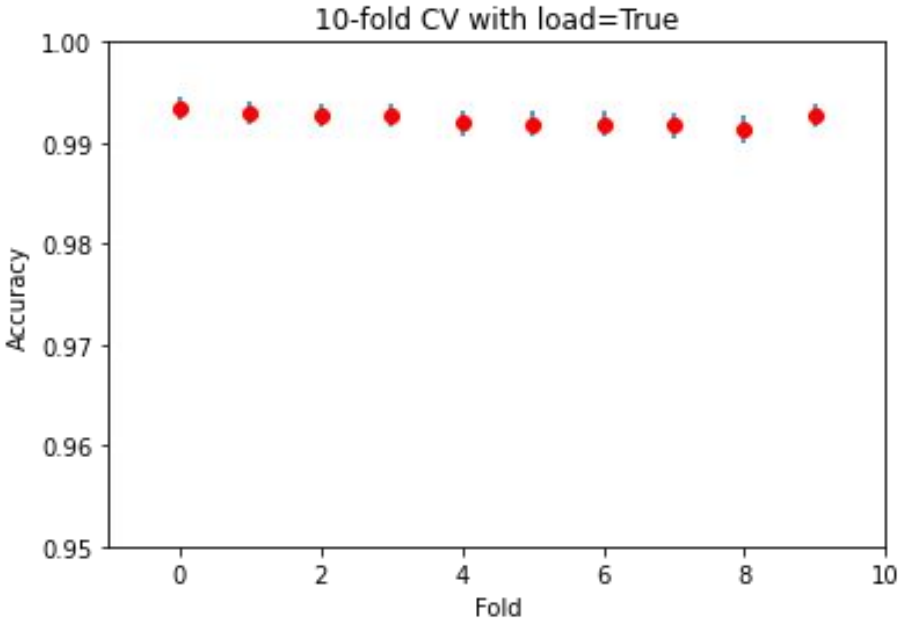
4. Evaluación del resto de etiquetadores (III)

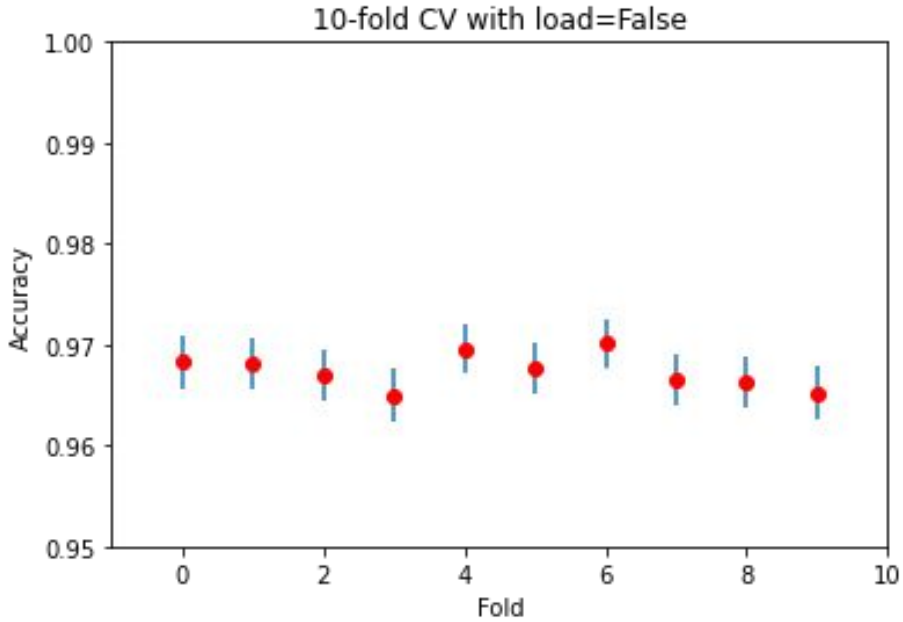
CRF Tagger



4. Evaluación del resto de etiquetadores (IV)

Perceptron





5. Evaluación del paquete Freeling

1. **Documentación**
   Descriptiva pero dispersa. Demasiadas redirecciones.

2. **Facilidad de instalación**
   a. Ubuntu 20.04 ✘
   b. Windows 10 pirata ✘
   c. Windows 10 ★

3. **Facilidad de uso**
   freeling\bin\analyze.bat -f es.cfg < Alicia_utf8.txt > .\results.txt

4. **Funcionalidad**
   Obtención de fichero txt con la etiqueta morfosintáctica correspondiente.

A_través_de/SP
la/DA0FS0
tarde/NCFS000

6. Conclusiones

- Experimentación con diferentes modelos (HMM, TNT, Brill Tagger, CRF o Perceptrón)
- Trabajar con menos etiquetas obtiene mejores resultados.
- Cuanta más información le pasemos al modelo, mejores prestaciones.