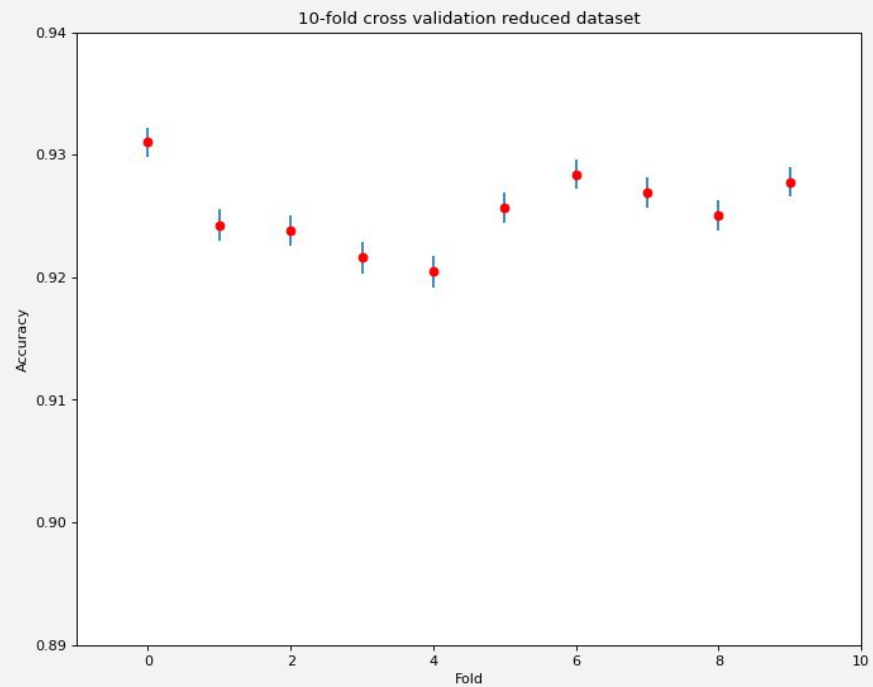
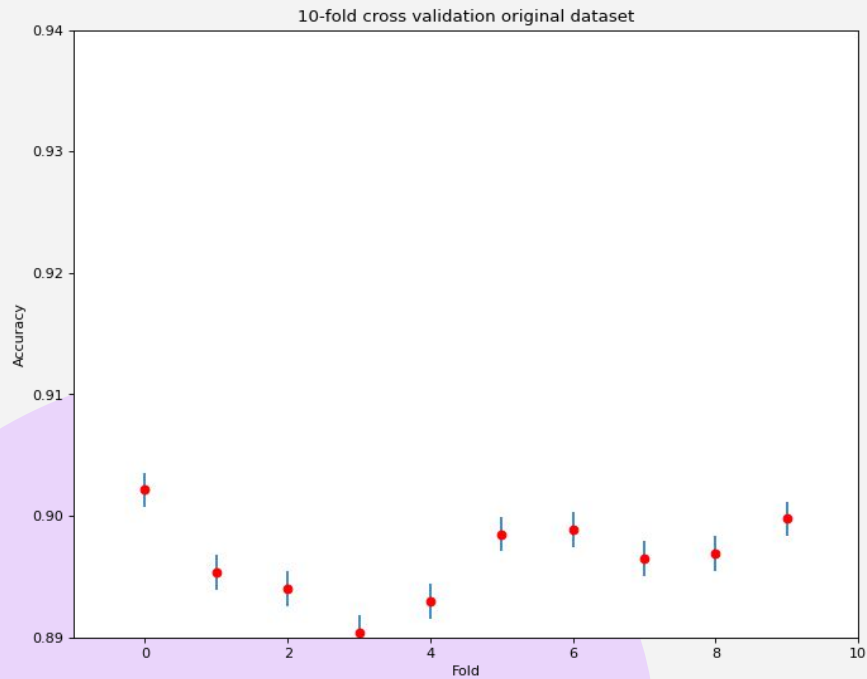


UNIVERSITAT POLITÈCNICA
DE VALÈNCIA

Evaluación de etiquetadores morfosintácticos

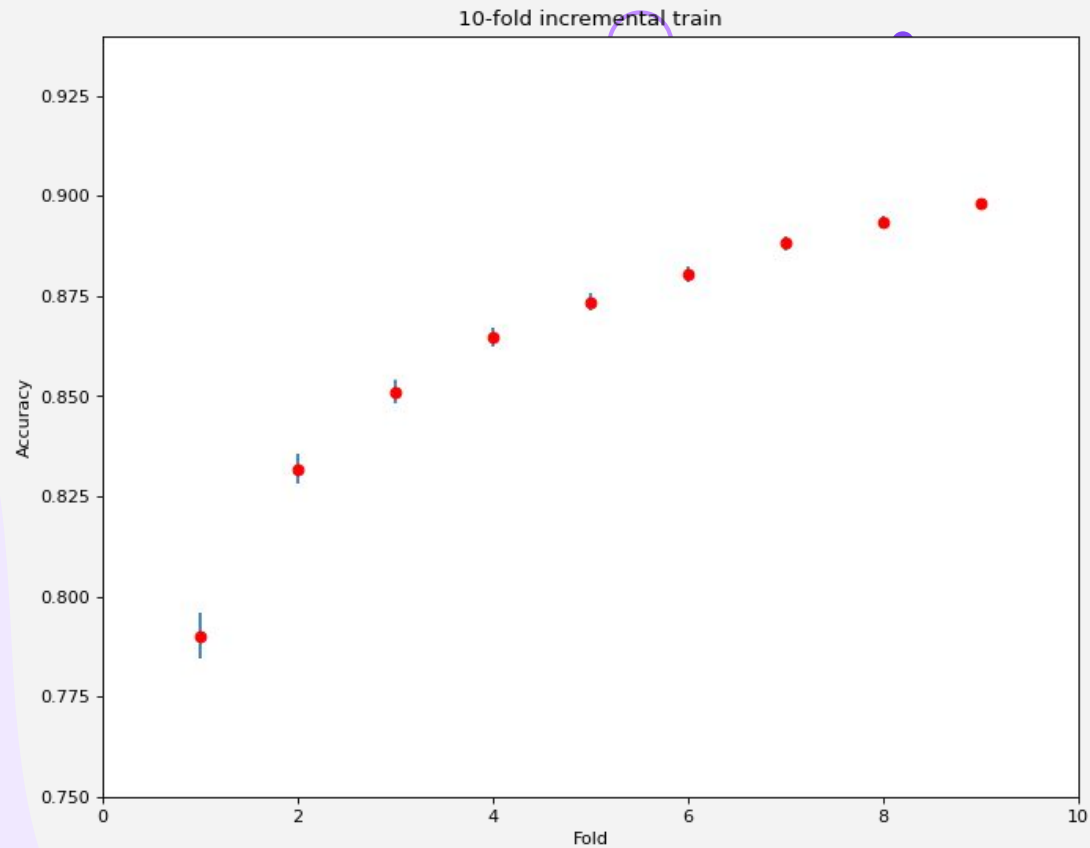
Lingüística Computacional

Luis Cardoza Bird



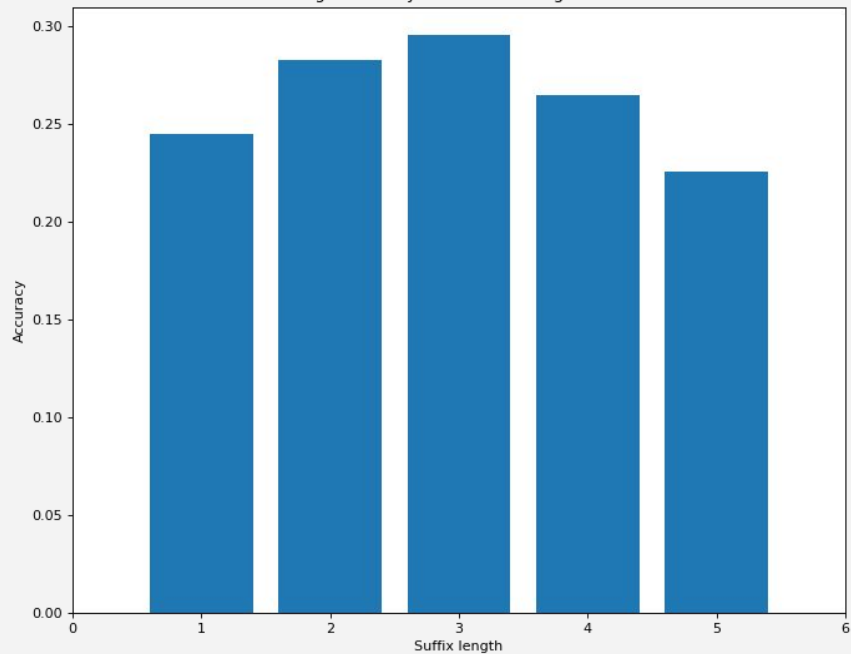
Tarea 1

Tarea 2

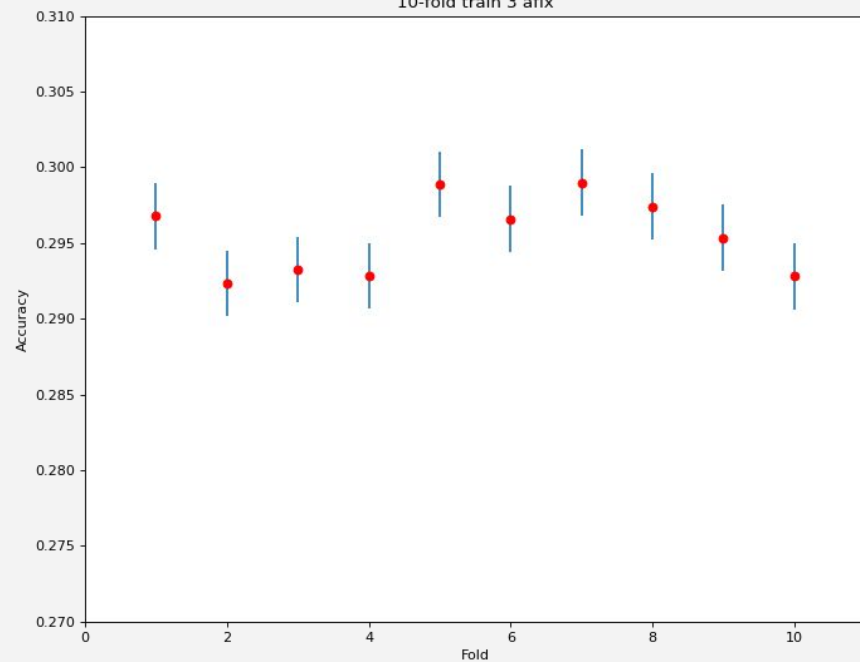


Tarea 3

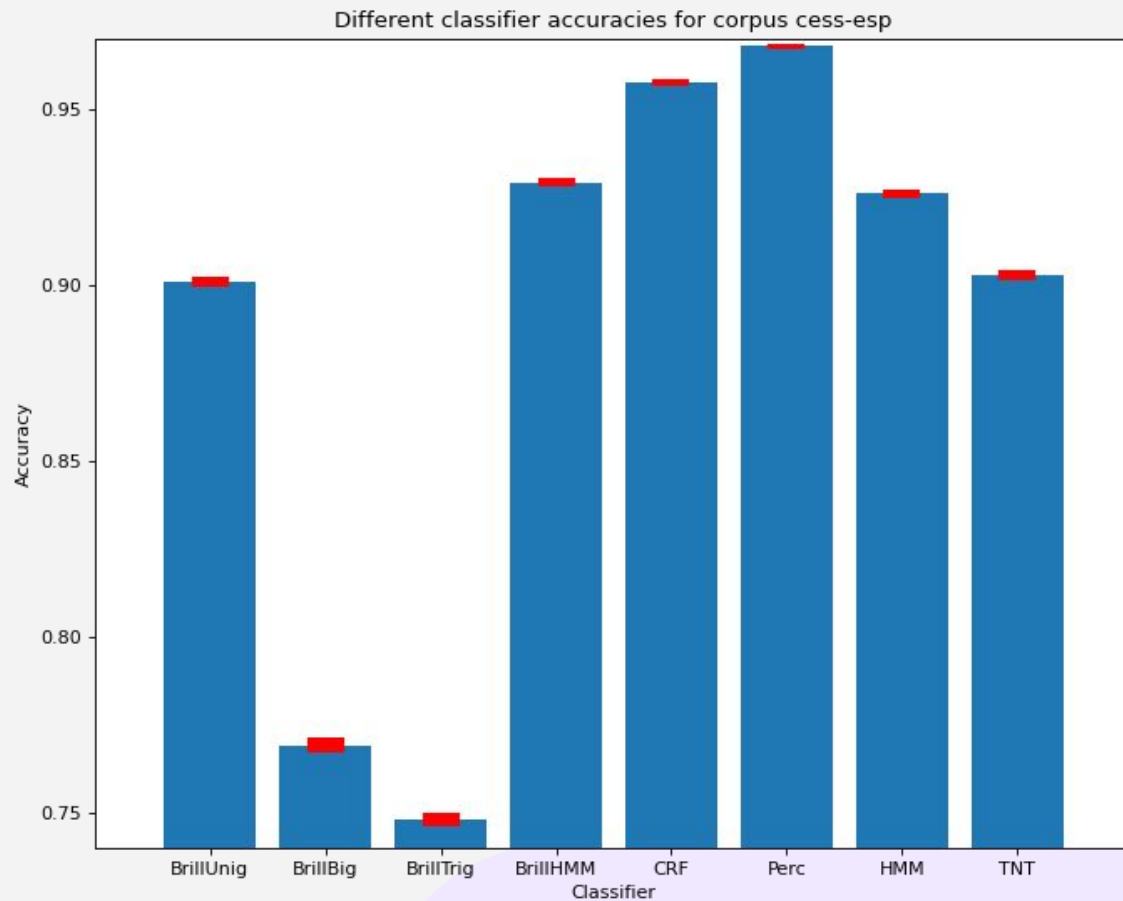
Average accuracy for different length suffixes



10-fold train 3 afix



Tarea 4



Tarea 5 Spacy



```
crdzbird@Luis-MacBook-Pro proyecto_final % python -m spacy download es_core_news_sm
```

```
You can now load the package via spacy.load('es_core_news_sm')
```

```
nteligenciaArtificialUPV/LC/proyecto_final/spaCy.py
```

```
SPACE  
A ADP  
través NOUN  
de ADP  
la DET  
tarde NOUN  
color NOUN  
de ADP  
oro NOUN
```

```
SPACE  
el DET  
agua NOUN  
nos PRON  
lleva VERB  
sin ADP  
esfuerzo NOUN  
por ADP  
nuestra DET  
parte NOUN  
, PUNCT
```

```
SPACE  
pues SCONJ  
los DET  
que PRON  
empujan VERB  
los DET  
remos VERB
```

```
SPACE  
son AUX
```

```
Requirement already satisfied: mpmath==0.19 in /opt/homebrew/lib/python3.11/site-packages (from sympy->torch>=1.3.0->stanza) (1.3.0)
Downloading stanza-1.6.1-py3-none-any.whl (881 kB)
881.2/881.2 kB 15.4 MB/s eta 0:00:00
Downloading emoji-2.8.0-py2.py3-none-any.whl (358 kB)
358.9/358.9 kB 23.2 MB/s eta 0:00:00
Installing collected packages: emoji, stanza
Successfully installed emoji-2.8.0 stanza-1.6.1

[notice] A new release of pip is available: 23.2.1 -> 23.3.1
[notice] To update, run: python3.11 -m pip install --upgrade pip
crdzbird@Luis-MacBook-Pro proyecto_final % cd /Users/crdzbird/Documents/MasterInteligenciaArtificialUPV/LC/proyecto_final ; /usr/bin/env /opt/homebrew/bin/python3 /Users/crdzbird/.vscode/extensions/ms-python.python-2023.18.0/pythonFiles/lib/python/buggy/adapters/../../buggy/launcher 63757 -- /Users/crdzbird/Documents/MasterInteligenciaArtificialUPV/LC/proyecto_final/stanza.py
crdzbird@Luis-MacBook-Pro proyecto_final % cd /Users/crdzbird/Documents/MasterInteligenciaArtificialUPV/LC/proyecto_final ; /usr/bin/env /opt/homebrew/bin/python3 /Users/crdzbird/.vscode/extensions/ms-python.python-2023.18.0/pythonFiles/lib/python/buggy/adapters/../../buggy/launcher 63774 -- /Users/crdzbird/Documents/MasterInteligenciaArtificialUPV/LC/proyecto_final/stanza.py
2023-10-29 09:38:18 INFO: Checking for updates to resources.json in case models have been updated. Note: this behavior can be turned off with the --no-update flag.
d.REUSE_RESOURCES
Downloading https://raw.githubusercontent.com/stanfordnlp/stanza-resources/main/resources_1.6.0.json: 367kB [00:00, 156MB/s]
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/tokenize/ancora.pt: 100%| 636
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/mwt/ancora.pt: 100%| 601k/601
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/pos/ancora_charlm.pt: 100%| 3
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/lemma/ancora_nocharlm.pt: 100%|
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/constituency/combined_charlm.pt
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/depparse/ancora_charlm.pt: 100%
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/sentiment/tass2020.pt: 100%|
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/ner/conll02.pt: 100%| 67.0M/6
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/pretrain/fasttextwiki.pt: 100%|
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/forward_charlm/newswiki.pt: 100
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/pretrain/conll17.pt: 100%| 10
Downloading https://huggingface.co/stanfordnlp/stanza-es/resolve/v1.6.0/models/backward_charlm/newswiki.pt: 10
2023-10-29 09:38:59 INFO: Loading these models for language: es (Spanish):
```

Processor	Package
tokenize	ancora
mwt	ancora
pos	ancora_charlm
lemma	ancora_nocharlm
constituency	combined_charlm
depparse	ancora_charlm
sentiment	tass2020
ner	conll02

```
2023-10-29 09:38:59 INFO: Using device: cpu
2023-10-29 09:38:59 INFO: Loading: tokenize
2023-10-29 09:38:59 INFO: Loading: mwt
2023-10-29 09:38:59 INFO: Loading: pos
2023-10-29 09:38:59 INFO: Loading: lemma
2023-10-29 09:38:59 INFO: Loading: constituency
2023-10-29 09:39:00 INFO: Loading: depparse
2023-10-29 09:39:00 INFO: Loading: sentiment
2023-10-29 09:39:00 INFO: Loading: ner
2023-10-29 09:39:01 INFO: Done loading processors!
crdzbird@Luis-MacBook-Pro proyecto_final %
```

Tarea 5 Stanza

1			SPACE
2	A	ADP	
3	través	NOUN	
4	de	ADP	
5	la	DET	
6	tarde	NOUN	
7	color	NOUN	
8	de	ADP	
9	oro	NOUN	
10			
11			SPACE
12	el	DET	
13	agua	NOUN	
14	nos	PRON	
15	lleva	VERB	
16	sin	ADP	
17	esfuerzo	NOUN	
18	por	ADP	
19	nuestra	DET	
20	parte	NOUN	
21	,	PUNCT	
22			
23			SPACE
24	pues	SCONJ	
25	los	DET	
26	que	PRON	
27	empujan	VERB	
28	los	DET	
29	remos	VERB	
30			
31			SPACE
32	son	AUX	
33	unos	DET	
34	brazos	NOUN	
35	infantiles	ADJ	
36			
37			SPACE
38	que	PRON	
39	intentan	VERB	
40	,	PUNCT	
41	con	ADP	
42	sus	DET	
43	manitas	NOUN	
44			
45			SPACE
46	guiar	VERB	
47	el	DET	
48	curso	NOUN	
49	de	ADP	
50	nuestra	DET	
51	barca	NOUN	
52	.	PUNCT	
53			
54			
55			SPACE

	1	A	ADP
	2	través	NOUN
	3	de	ADP
	4	la	DET
	5	tarde	NOUN
	6	color	NOUN
	7	de	ADP
	8	oro	NOUN
	9	el	DET
	10	agua	NOUN
	11	nos	PRON
	12	lleva	VERB
	13	sin	ADP
	14	esfuerzo	NOUN
	15	por	ADP
	16	nuestra	DET
	17	parte	NOUN
	18	,	PUNCT
	19	pues	SCONJ
	20	los	DET
	21	que	PRON
	22	empujan	VERB
	23	los	DET
	24	remos	NOUN
	25	son	AUX
	26	unos	DET
	27	brazos	NOUN
	28	infantiles	ADJ
	29	que	PRON
	30	intentan	VERB
	31	,	PUNCT
	32	con	ADP
	33	sus	DET
	34	manitas	NOUN
	35	guiar	VERB
	36	el	DET
	37	curso	NOUN
	38	de	ADP
	39	nuestra	DET
	40	barca	NOUN
	41	.	PUNCT
	42	Pero	CCONJ
	43	,	PUNCT
	44	i	PUNCT
	45	las	DET
	46	tres	NUM
	47	son	AUX
	48	muy	ADV
	49	crueles	ADJ
	50	!	PUNCT
	51	ya	ADV
	52	que	SCONJ
	53	sin	ADP
	54	fijar	VERB
	55	se	PRON



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



```
# Importar las bibliotecas necesarias
import nlk      Import "nlk" could not be resolved
import json
import spacy

from nlk.corpus import cess_esp    Import "nlk.corpus" could not be resolved
from spacy.tokens import Doc
from spacy.training import Example
!pip install es_core_news_sm      Use '%pip install' instead of '!pip install'

# Descargar el corpus CESS-ESP
nlk.download('cess_esp')
```

[2]






```
# Obtener las oraciones etiquetadas
sentences = cess_esp.tagged_sents()

# Convertir las oraciones al formato de spaCy y dividir en entrenamiento y prueba
def convert_to_spacy_format(sentences):
    data = []
    for sent in sentences:
        words, tags = zip(*sent)
        data.append(" ".join(words), {"words": words, "tags": tags})
    return data

all_data = convert_to_spacy_format(sentences)

# Dividir en entrenamiento y prueba
split = int(0.8 * len(all_data))
train_data = all_data[:split]
test_data = all_data[split:]
```



```
# Cargar el modelo base en español y preparar los datos
nlp = spacy.blank("es")

# Añadir el etiquetador al pipeline
nlp.add_pipe("tagger")

# Añadir las etiquetas al etiquetador
tagger = nlp.get_pipe("tagger")
for _, annotations in train_data:
    for tag in annotations.get("tags"):
        tagger.add_label(tag)

# Entrenar el modelo
optimizer = nlp.begin_training()
for itn in range(10):
    for text, annotations in train_data:
        doc = nlp.make_doc(text)
        example = Example.from_dict(doc, annotations)
        nlp.update([example], drop=0.2, sgd=optimizer)
```

```
# Evaluar el modelo en el conjunto de prueba
correct = 0
total = 0
for text, annotations in test_data:
    doc = nlp(text)
    predicted_tags = [token.tag_ for token in doc]
    true_tags = annotations["tags"]
    correct += sum(1 for p, t in zip(predicted_tags, true_tags) if p == t)
    total += len(true_tags)

accuracy = correct / total
print(f"Accuracy: {accuracy:.4f}")
```

Accuracy: 0.6009