Informe: Extracció d'entitats anomenades (pràctica 3)

Pol Rion Solè
Gerard Gómez Izquierdo
GIA-PLH
16/5/2023

ÍNDEX

Primera part

Introducció

En l'àmbit del processament del llenguatge natural, l'extracció de les entitats anomenades és una tasca fonamental que implica identificar i classificar les entitats específiques en un text, com ara noms de persones, organitzacions i llocs geogràfics entre d'altres.

En aquest treball, ens centrem en la implementació de dos sistemes NER, un pel Castellà i l'altre per l'Holandès. Per aconseguir-ho farem ús de Conditional Random Fields (CRF). Els CRF són un model probabilístic que és àmpliament utilitzat en tasques d'etiquetatge d'entitats anomenades, el qual destaca per la seva capacitat de tenir en compte simultàniament tant característiques locals com del context per millorar la precisió de la identificació d'entitats de termes.

El nostre objectiu és desenvolupar dos sistemes NER precisos mitjançant l'ús de CRF. Per fer-ho, explorarem diverses característiques del text com la informació lèxica, morfològica i sintàctica entre d'altres i provarem diferents maneres de codificar el corpus d'entrenament, entre les quals es troben les codificacions BIO, IO i BIOW.

Corpus

Per entrenar i avaluar els nostres models utilitzarem el corpus *conll2002* que ens proporciona la llibreria de python NLTK. Aquest corpus consisteix en textos de notícies de periòdics en espanyol i neerlandès, que han estat anotats manualment amb informació sobre les entitats anomenades que apareixen al text.

En concret, el corpus *conll2002* inclou un total de 11755 frases en espanyol i 23896 frases en neerlandés, que s'han anotat amb el postag dels termes i la informació sobre les entitats de les categories "PER" (persones), "ORG" (organitzacions), "LOC" (llocs geogràfics) i “MISC” (altres entitats). Així, cada paraula del text ha estat anotada amb una etiqueta que indica a quina entitat pertany.

Aquest corpus ve separat en 3 particions: train, test_a i test_b. Durant el treball farem servir la partició de train per entrenar el model, la de test_a com a validació i la de test_b com a test.

Creació del CRF

Abans de començar a explicar les diferents característiques que hem anat utilitzant per millorar l'entrenament del nostre CRF volem destacar com hem implementat aquest model i quines característiques del text mirem per defecte.

En un inici vam usar la classe de nltk CRFTagger(), però ens veiem limitats a l'hora d'utilitzar certes funcions característiques, com per exemple la funció característica que té en compte el postag de la paraula. Això succeïa perquè per accedir al postag de cada terme hi havia

dues opcions: calcular el postag dins de la funció get_features , la qual cosa implica un cost molt gran en l'àmbit computacional i temporal, ja que es crida per cada token o passar-li com a paràmetre el postag de cada paraula a la classe CRFTagger, la qual cosa no era viable, ja que la classe està implementada per rebre la informació dels termes amb la forma (word, label) i no era viable afegir un tercer element a no ser que es modifiqués la implementació de la classe.

Per aquesta raó vàrem decidir fer servir el CRF de sklearn, perquè aquest sí que ens permet passar-li com a paràmetres el postag i altres característiques que nosaltres vulguem de les paraules, de manera que dins de la funció equivalent a get_features només necessitarem afegir la característica i no caldrà calcular-la durant l'entrenament ni la predicció.

Per partir des del mateix punt inicial que amb el CRFTagger de NLTK, hem establert les mateixes funcions característiques per defecte que té aquesta classe, les quals són: Paraula actual, És majúscula?, Té signes de puntuació?, Té números?, i Sufixos de fins a mida 3.

```python
def word2features(sent, i):
    word = sent[i][0]
    features = {
        'word': word,
        'word.isdigit()': word.isdigit(),
        'punctuation': any(p in word for p in punctuation),
        'word[-3:]': word[-3:],
        'word[-2:]': word[-2:],
        'word[-1:]': word[-1:],
        'word.isupper()': word.isupper(),
        'word.istitle()': word.istitle(),
    }
    return features
```

Figura 1: Funció equivalent a get_features que estableix les funcions característiques amb les que entrrenarem el CRF

Feature Functions

En aquest apartat del treball anirem veient com afecten les diferents funcions característiques a les prediccions del nostre model sobre la partició de test_a, la qual utilitzarem com a partició de validació. Per cada funció que afegim, compararem el resultat amb l'anterior i si millora encara treballarem amb aquesta. Finalment, obtindrem un model CRF entrenat per cada idioma amb aquelles feature functions que milloren la precisió i el recall del CRF en la partició de validació de l'idioma corresponent.

Per a avaluar el nostre model, prestarem atenció a la mètrica f1-score macro. Aquesta decisió ha estat presa, ja que és una mètrica que no es veu afectada per el desbalanceig de les classes i té en compte la precisió i el recall. Com que nosaltres hem codificat el nostre text de manera que les paraules d'aquest pertanyin a una determinada classe, hi ha classes que es veuen altament sobrerepresentades per aquesta classificació. Per exemple, en la codificació iob, el nombre de paraules etiquetades amb 'O' predomina sobre totes les altres classes amb diferència. D'aquesta manera, doncs, donem la mateixa importància a que el model predigui lo millor possible totes les classes de manera equitativa i que per exemple, al

millorar el model afegint feature functions, escollim aquelles que milloren totes les classes i no només aquelles que més cops apareixen en el corpus.

Cal considerar també que per tal de dur el model a les seves millors condicions, primer farem la selecció de característiques sobre una codificació fixada, concretament la BIO. Un cop aquestes estiguin decidides i seleccionades, provarem amb altres codificacions diferents. Posteriorment, decodificarem els resultats predits amb les diferents codificacions i compararem els resultats per tal de veure quina d'elles és més favorable al model.

Finalment, cal destacar que per analitzar el model, cada cop que afegim una feature function ens fixarem en els resultats de les prediccions sense decodificar sobre la partició test_a, la qual utilitzarem com a validació. Això ho fem perquè volem una avaluació més precisa. Al utilitzar les etiquetes B i I en el classification report, estem tenint en compte el rendiment entre les etiquetes d'inici i les internes de les entitats. Aquest fet pot revelar informació important sobre la capacitat del model per identificar correctament el començament d'una entitat (etiqueta B) i per seguir correctament la seva continuació (etiqueta I). Si analitzessim només les entitats sense la seva codificació, perdriem aquesta distinció i no obtindriem una avaluació precisa del model. Això podria portar a afegir feature functions que no ajuden al model a diferenciar entre termes d'inici o interns sobre l'entitat i que, per tant, al predir altres corpus de test es falles més quan ens trobem amb entitats compostes per diferents paraules.

Model Castellà

Features per defecte

A continuació es poden apreciar els resultats amb amb les funcions característiques per defecte sobre test_a (validació):

	precision	recall	f1-score	support
B-LOC	0.65	0.78	0.71	984
B-MISC	0.58	0.49	0.53	445
B-ORG	0.82	0.72	0.77	1700
B-PER	0.86	0.75	0.80	1222
I-LOC	0.66	0.74	0.70	337
I-MISC	0.37	0.52	0.43	654
I-ORG	0.77	0.70	0.73	1366
I-PER	0.88	0.90	0.89	859
O	0.99	0.99	0.99	45356
accuracy			0.95	52923
macro avg	0.73	0.73	0.73	52923
weighted avg	0.95	0.95	0.95	52923

F1 score macro avg: 0.7277484385342833

Aquests resultats obtinguts ens serviran de baseline per a la construcció i l'ajust del nostre model. Es pot apreciar que la classe amb millors resultats és la O, ja que és la categoria

predominant en la codificació del text. Per altra banda, cal ressaltar que la mètrica global f1-score macro de la que partim serà de 0.727.

Pos-Tags

Després d'afegir al que ja teníem, la consideració de Pos-Tags a l'hora de predir les categories de les paraules, hem obtingut el següent:

	precision	recall	f1-score	support
B-LOC	0.64	0.78	0.71	984
B-MISC	0.59	0.48	0.53	445
B-ORG	0.81	0.73	0.77	1700
B-PER	0.84	0.75	0.79	1222
I-LOC	0.67	0.70	0.68	337
I-MISC	0.41	0.50	0.45	654
I-ORG	0.76	0.72	0.74	1366
I-PER	0.86	0.91	0.89	859
O	0.99	0.99	0.99	45356
accuracy			0.95	52923
macro avg	0.73	0.73	0.73	52923
weighted avg	0.95	0.95	0.95	52923

F1 score macro avg: 0.7282294957470687

S'ha pogut apreciar una lleugera millora en el resultat global de les prediccions, com bé mostra la mètrica avaluada. Algunes classes han presentat empitjorament en el seu f1-score, però d'altres l'han millorat. Com que aquesta característica ha fet precisar millor en les prediccions del model, l'afegirem a la llista de característiques a tenir en compte i prosseguirem amb les proves.

Lemmas

Considerant ara addicionalment els lemes de les paraules del text com a nova característica a l'hora de dur a terme les prediccions hem obtingut:

	precision	recall	f1-score	support
B-LOC	0.64	0.79	0.71	984
B-MISC	0.62	0.50	0.55	445
B-ORG	0.82	0.73	0.77	1700
B-PER	0.85	0.76	0.80	1222
I-LOC	0.65	0.77	0.70	337
I-MISC	0.42	0.52	0.46	654
I-ORG	0.78	0.70	0.74	1366
I-PER	0.86	0.92	0.89	859
O	0.99	0.99	0.99	45356
accuracy			0.95	52923
macro avg	0.74	0.74	0.74	52923
weighted avg	0.96	0.95	0.95	52923

F1 score macro avg: 0.7353427463689077

Es pot veure que la mètrica segueix creixent, per tant, seguim millorant. Algunes de les classes beneficiades per la inclusió d'aquesta nova característica han estat B-MISC, B-PER, I-LOC, I-MISC entre d'altres. El fet de considerar el lema de cada paraula a l'hora de predir ha ajudat al nostre model a afinar en les prediccions, identificant millor la classe de les paraules. Per aquest motiu, inclourem aquesta característica a la llista de característiques que valorem de les paraules i prosseguirem amb més proves.

Longitud

Afegint la longitud de la paraula com a característica, hem obtingut el següent:

	precision	recall	f1-score	support
B-LOC	0.64	0.78	0.70	984
B-MISC	0.64	0.49	0.56	445
B-ORG	0.83	0.73	0.77	1700
B-PER	0.83	0.77	0.80	1222
I-LOC	0.68	0.74	0.71	337
I-MISC	0.44	0.49	0.46	654
I-ORG	0.78	0.72	0.75	1366
I-PER	0.83	0.92	0.87	859
O	0.99	0.99	0.99	45356
accuracy			0.95	52923
macro avg	0.74	0.74	0.74	52923
weighted avg	0.95	0.95	0.95	52923

F1 score macro avg: 0.7351121599150342

Com es pot veure en aquest cas, el fet de considerar la longitud de les paraules com a característica addicional ha confós el nostre model i ha provocat que obtinguem uns pitjors resultats. Per aquest motiu, no la inclourem en el llistat de característiques i continuarem fent proves sense tenir-la en compte.

Prefixos

En aquesta ocasió hem afegit els prefixos de longitud fins a, com a màxim, 3 lletres i hem valorat com ha funcionat el model en aquestes condicions. Per a que s'entengui millor, el que hem fet ha estat, per cada paraula, afegir el prefix format per la seva lletra inicial, el prefix format per les seves dues lletres inicials i el prefix format per les seves tres lletres inicials. Tot això tenint en compte la llargada de la paraula i evitant casos contradictoris com podria ser, per exemple, afegir un prefix de llargada 3 en una paraula de llargada 2. Els resultats han estat els següents:

	precision	recall	f1-score	support
B-LOC	0.64	0.80	0.71	984
B-MISC	0.62	0.50	0.55	445
B-ORG	0.83	0.74	0.78	1700
B-PER	0.84	0.77	0.80	1222
I-LOC	0.65	0.74	0.69	337
I-MISC	0.45	0.54	0.49	654
I-ORG	0.78	0.69	0.73	1366
I-PER	0.83	0.92	0.87	859
O	0.99	0.99	0.99	45356
accuracy			0.95	52923
macro avg	0.74	0.74	0.74	52923
weighted avg	0.96	0.95	0.95	52923

F1 score macro avg: 0.7355328736477399

Es pot veure que hem millorat la mètrica considerada. És per aquest motiu que la inclourem en el nostre llistat de característiques a tenir en compte a l'hora de predir les classes de les paraules del text.

Arrels

Hem considerat també la possibilitat de treballar amb les arrels de les paraules trobades al text, són una tècnica similar a la de la consideració de lemes, però hem pensat que potser podia servir com a informació addicional favorable per al treball predictiu del nostre model. S'ha obtingut el següent:

	precision	recall	f1-score	support
B-LOC	0.63	0.80	0.70	984
B-MISC	0.62	0.51	0.56	445
B-ORG	0.83	0.73	0.78	1700
B-PER	0.85	0.77	0.80	1222
I-LOC	0.64	0.74	0.69	337
I-MISC	0.44	0.50	0.47	654
I-ORG	0.77	0.69	0.73	1366
I-PER	0.82	0.92	0.87	859
O	0.99	0.99	0.99	45356
accuracy			0.95	52923
macro avg	0.73	0.74	0.73	52923
weighted avg	0.96	0.95	0.95	52923

F1 score macro avg: 0.732309189513603

Es pot veure que el que hem provocat és sobreinformar el nostre model, provocant així que cometi més errors a l'hora de predir. Així doncs, els resultats obtinguts han estat desfavorables. Per aquest motiu, les arrels de les paraules no seran considerades com a característica per a entrenar el model.

Gazetteers

Els gazetteers són llistes de paraules relacionades en un ambit. Això el que ens permet es trobar si una paraula pertany a aquella llista i afegir-li una etiqueta relacionada amb aquesta perquè el CRF tingui més informació sobre aquesta a l'hora de predir la seva entitat.

En el nostre cas hem decidit crear 4 gazetteers, un per cada etiqueta del nostre NER. Aquestes llistes les hem creat amb la partició de train, aquelles paraules que tenen B- algo o I-algo les hem afegit a la llista que els hi correspon de les 4. Per tant, ara si el model es troba amb una d'aquestes paraules li afegirà la característica localització, organització, persona o miscellaneous.

Provem d'afegir els gazetteers un a un per veure quin impacte tenen en els resultats del nostre model.

1) Gazeteer de localització

	precision	recall	f1-score	support
B-LOC	0.69	0.78	0.73	984
B-MISC	0.56	0.51	0.53	445
B-ORG	0.85	0.63	0.72	1700
B-PER	0.91	0.55	0.68	1222
I-LOC	0.49	0.70	0.58	337
I-MISC	0.57	0.51	0.54	654
I-ORG	0.81	0.60	0.69	1366
I-PER	0.94	0.68	0.79	859
O	0.97	1.00	0.98	45356
accuracy			0.94	52923
macro avg	0.75	0.66	0.69	52923
weighted avg	0.94	0.94	0.94	52923

F1 score macro avg: 0.6938410520369449

2) Gazeteer d'organització

	precision	recall	f1-score	support
B-LOC	0.62	0.79	0.69	984
B-MISC	0.50	0.54	0.52	445
B-ORG	0.88	0.60	0.71	1700
B-PER	0.79	0.82	0.80	1222
I-LOC	0.52	0.72	0.61	337
I-MISC	0.37	0.57	0.44	654
I-ORG	0.84	0.49	0.62	1366
I-PER	0.81	0.93	0.86	859
O	0.99	0.99	0.99	45356
accuracy			0.95	52923
macro avg	0.70	0.72	0.69	52923
weighted avg	0.95	0.95	0.95	52923

F1 score macro avg: 0.6946604256229263

3) Gazetteer de persona

	precision	recall	f1-score	support
B-LOC	0.53	0.80	0.64	984
B-MISC	0.55	0.52	0.54	445
B-ORG	0.76	0.74	0.75	1700
B-PER	0.95	0.50	0.66	1222
I-LOC	0.46	0.73	0.56	337
I-MISC	0.41	0.55	0.47	654
I-ORG	0.72	0.72	0.72	1366
I-PER	0.96	0.61	0.74	859
O	0.99	0.99	0.99	45356
accuracy			0.94	52923
macro avg	0.70	0.69	0.67	52923
weighted avg	0.95	0.94	0.94	52923

F1 score macro avg: 0.6748057127389856

4) Gazetteer de miscellaneous

	precision	recall	f1-score	support
B-LOC	0.63	0.79	0.70	984
B-MISC	0.74	0.44	0.55	445
B-ORG	0.80	0.74	0.77	1700
B-PER	0.83	0.78	0.81	1222
I-LOC	0.59	0.73	0.66	337
I-MISC	0.75	0.37	0.49	654
I-ORG	0.73	0.74	0.73	1366
I-PER	0.82	0.93	0.87	859
O	0.99	0.99	0.99	45356
accuracy			0.96	52923
macro avg	0.76	0.72	0.73	52923
weighted avg	0.96	0.96	0.95	52923

F1 score macro avg: 0.7306189940875643

Veiem que després d'haver tots els gazetteers un a un, cap d'ells proporciona uns resultats millors als que ja havíem obtingut prèviament. Per tant, ometrem la inclusió de gazetteers en la nostra consideració de característiques.

Context de la paraula

En darrer lloc, s'ha plantejat que les features del context que envolta la paraula valorada pot aportar una informació interessant i útil a l'hora de predir la etiqueta d'aquesta mencionada paraula. La implementació d'aquesta característica ha consistit en considerar les característiques de la paraula anterior i posterior a la paraula valorada en el moment.

	precision	recall	f1-score	support
B-LOC	0.61	0.78	0.68	984
B-MISC	0.61	0.51	0.55	445
B-ORG	0.81	0.73	0.77	1700
B-PER	0.86	0.76	0.80	1222
I-LOC	0.57	0.75	0.65	337
I-MISC	0.50	0.53	0.51	654
I-ORG	0.79	0.69	0.73	1366
I-PER	0.86	0.91	0.88	859
O	0.99	0.99	0.99	45356
accuracy			0.95	52923
macro avg	0.73	0.74	0.73	52923
weighted avg	0.96	0.95	0.95	52923

F1 score macro avg: 0.7310415561180614

Es pot veure que, tot i que en un primer moment semblava una bona idea, aquesta característica no ha aportat els resultats esperats. No hem arribat a superar la mètrica de f1 score més elevada que teníem fins el moment i, per tant, aquesta característica no la tindrem en compte en el nostre llistat de característiques.

Podem concloure, doncs, que el nostre llistat de característiques per a realitzar la predicció de l'etiqueta de les paraules del text en espanyol amb el que treballem consta de:

- Característiques per defecte
- Pos-Tag
- Lemmas
- Prefixos

Ara després d'haver ajustat el model anem a desfer la codificació BIO de les prediccions i anem a veure quin ha estat el millor resultat predint les entitats de les parules sense la 'B' i la 'I'.

	precision	recall	f1-score	support
LOC	0.65	0.80	0.72	1321
MISC	0.52	0.54	0.53	1099
O	0.99	0.99	0.99	45356
ORG	0.82	0.73	0.77	3066
PER	0.84	0.83	0.84	2081
accuracy			0.96	52923
macro avg	0.76	0.78	0.77	52923
weighted avg	0.96	0.96	0.96	52923

F1 score macro avg: 0.7686763504778588

Model Neerlandés

Seguim el mateix procediment que anteriorment.

Features per defecte

Per començar, anem a analitzar els resultats que obtenim utilitzant les funcions característiques per defecte sobre test_a (validació):

	precision	recall	f1-score	support
B-LOC	0.68	0.61	0.64	479
B-MISC	0.69	0.66	0.67	748
B-ORG	0.88	0.59	0.71	686
B-PER	0.64	0.74	0.69	703
I-LOC	0.47	0.28	0.35	64
I-MISC	0.27	0.48	0.34	215
I-ORG	0.76	0.63	0.69	396
I-PER	0.75	0.91	0.82	423
O	0.99	0.99	0.99	33973
accuracy			0.96	37687
macro avg	0.68	0.65	0.66	37687
weighted avg	0.96	0.96	0.96	37687

F1 score macro avg: 0.6562139859287889

Com es pot veure en el classification report superior obtenim un model baseline amb un f1 score de 0.65621. Aquest valor ens servirà com a base per decidir si afegim la següent feature function. Si a l'afegir una nova feature millorem, ens quedarem amb aquesta i continuarem treballant sobre les anteriors més la que ha millorat el resultat, si no no l'afegirem al nostre model i encara provarem de noves.

POS-Tags

La primera característica que provarem d'afegir en el nostre CRF és el postag de cada paraula. Per fer-ho l'hem passat com a paràmetre i l'hem afegit a les característiques de la paraula a la funció word2features.

	precision	recall	f1-score	support
B-LOC	0.67	0.62	0.65	479
B-MISC	0.68	0.68	0.68	748
B-ORG	0.86	0.56	0.68	686
B-PER	0.64	0.74	0.68	703
I-LOC	0.43	0.31	0.36	64
I-MISC	0.27	0.53	0.36	215
I-ORG	0.75	0.62	0.68	396
I-PER	0.75	0.91	0.82	423
O	0.99	0.99	0.99	33973
accuracy			0.96	37687
macro avg	0.67	0.66	0.66	37687
weighted avg	0.96	0.96	0.96	37687

F1 score macro avg: 0.6567949161665281

Com s'aprecia en els resultats hem millorat una mica la mètrica d'f1 score a 0.65679. Pel que fa a específicament cada etiqueta podem veure com hem millorat l'f1 score d'algunes com B-LOC o B-MISC i hem empitjorat unes altres com B-ORG d'entre altres. El balanç final de millora pèrdua és positiu, per tant, afegirem aquesta funció característica i continuarem treballant amb ella.

Lemmas

La següent característica que afegirem seran els lemmas de les paraules. Per fer-ho, els passarem com a paràmetres com hem fet amb els POS-tags, però en aquest cas hauran de ser prèviament calculats i afegits al corpus d'entrenament i de posterior predicció. Per calcular-los hem utilitzat SpaCy.

	precision	recall	f1-score	support
B-LOC	0.67	0.61	0.64	479
B-MISC	0.70	0.68	0.69	748
B-ORG	0.90	0.59	0.71	686
B-PER	0.62	0.73	0.67	703
I-LOC	0.42	0.30	0.35	64
I-MISC	0.35	0.53	0.42	215
I-ORG	0.82	0.65	0.72	396
I-PER	0.74	0.92	0.82	423
O	0.99	0.99	0.99	33973
accuracy			0.96	37687
macro avg	0.69	0.67	0.67	37687
weighted avg	0.96	0.96	0.96	37687

F1 score macro avg: 0.6679902468712986

Com podem veure en els resultats aconseguim millorar l'f1 score a 0.66799. Si ens fixem en les etiquetes, podem veure com l'f1 score de B-ORG, I-MISC i I-ORG millora notablement respecte a l'anterior prova feta amb POS-tags, per tant, afegirem també aquesta característica a la nostra funció word2features.

Longitud

A continuació provarem d'afegir la característica de la longitud de la paraula.

	precision	recall	f1-score	support
B-LOC	0.73	0.63	0.68	479
B-MISC	0.71	0.68	0.69	748
B-ORG	0.91	0.60	0.72	686
B-PER	0.64	0.78	0.70	703
I-LOC	0.49	0.30	0.37	64
I-MISC	0.35	0.52	0.42	215
I-ORG	0.84	0.65	0.73	396
I-PER	0.75	0.93	0.83	423
O	0.99	0.99	0.99	33973
accuracy			0.96	37687
macro avg	0.71	0.68	0.68	37687
weighted avg	0.96	0.96	0.96	37687

F1 score macro avg: 0.6824504327807428

Com es pot veure en el classification report millorem l'f1 score a 0.68245. Destaquem una millora notable en els f1 scores específics de B-PER i una millora lleu en la gran majoria d'ells. Com hem millorat l'f1 score macro avg també afegirem aquesta nova feature a la funció word2features.

Prefixos

En aquesta nova feature provarem d'afegir els prefixos de fins a mida 3 de les paraules.

	precision	recall	f1-score	support
B-LOC	0.72	0.66	0.69	479
B-MISC	0.75	0.74	0.75	748
B-ORG	0.91	0.63	0.74	686
B-PER	0.65	0.78	0.71	703
I-LOC	0.39	0.30	0.34	64
I-MISC	0.29	0.50	0.37	215
I-ORG	0.82	0.66	0.73	396
I-PER	0.76	0.90	0.82	423
O	0.99	0.99	0.99	33973
accuracy			0.96	37687
macro avg	0.70	0.68	0.68	37687
weighted avg	0.97	0.96	0.96	37687

F1 score macro avg: 0.6821312158846061

Tal i com podem veure el f1 score avg ha disminuït respecte el màxim que haviem torbat, per tant, descartarem afegir aquesta nova feature i seguirem provant de noves.

Arrel

Seguidament veurem com reacciona el nostre model al afegir una nova feature que contingui l'arrel de la paraula.

	precision	recall	f1-score	support
B-LOC	0.73	0.65	0.69	479
B-MISC	0.73	0.66	0.69	748
B-ORG	0.91	0.63	0.74	686
B-PER	0.63	0.77	0.69	703
I-LOC	0.45	0.30	0.36	64
I-MISC	0.35	0.49	0.41	215
I-ORG	0.84	0.64	0.73	396
I-PER	0.75	0.92	0.83	423
O	0.99	0.99	0.99	33973
accuracy			0.96	37687
macro avg	0.71	0.67	0.68	37687
weighted avg	0.96	0.96	0.96	37687

F1 score macro avg: 0.6808351525635238

Es pot apreciar com afegint l'arrel tampoc millorem el f1 score avg que haviem trobat fins ara, per tant la descartarem.

Gazetteers

A continuació provarem d'afegir els gazetteers ja esmentats

1) Gazetteer de localització

	precision	recall	f1-score	support
B-LOC	0.94	0.52	0.67	479
B-MISC	0.69	0.69	0.69	748
B-ORG	0.88	0.61	0.72	686
B-PER	0.58	0.80	0.67	703
I-LOC	0.69	0.28	0.40	64
I-MISC	0.28	0.49	0.36	215
I-ORG	0.80	0.65	0.72	396
I-PER	0.74	0.91	0.81	423
O	0.99	0.99	0.99	33973
accuracy			0.96	37687
macro avg	0.73	0.66	0.67	37687
weighted avg	0.96	0.96	0.96	37687

F1 score macro avg: 0.6696600152325431

2) Gazetteer d'organització

	precision	recall	f1-score	support
B-LOC	0.71	0.64	0.67	479
B-MISC	0.67	0.68	0.68	748
B-ORG	0.93	0.47	0.62	686
B-PER	0.60	0.78	0.68	703
I-LOC	0.40	0.31	0.35	64
I-MISC	0.33	0.52	0.40	215
I-ORG	0.98	0.45	0.61	396
I-PER	0.71	0.93	0.81	423
O	0.99	0.99	0.99	33973
accuracy			0.96	37687
macro avg	0.70	0.64	0.65	37687
weighted avg	0.96	0.96	0.96	37687

F1 score macro avg: 0.6455233229266519

3) Gazetteer de persona

	precision	recall	f1-score	support
B-LOC	0.52	0.78	0.62	479
B-MISC	0.66	0.68	0.67	748
B-ORG	0.79	0.64	0.71	686
B-PER	0.89	0.42	0.57	703
I-LOC	0.23	0.48	0.31	64
I-MISC	0.28	0.51	0.36	215
I-ORG	0.69	0.69	0.69	396
I-PER	0.89	0.46	0.61	423
O	0.99	0.99	0.99	33973
accuracy			0.95	37687
macro avg	0.66	0.63	0.62	37687
weighted avg	0.96	0.95	0.95	37687

F1 score macro avg: 0.6157812094693619

4) Gazetteer de miscellaneous

	precision	recall	f1-score	support
B-LOC	0.68	0.66	0.67	479
B-MISC	0.96	0.51	0.67	748
B-ORG	0.87	0.64	0.74	686
B-PER	0.60	0.80	0.69	703
I-LOC	0.45	0.33	0.38	64
I-MISC	0.93	0.30	0.45	215
I-ORG	0.76	0.68	0.72	396
I-PER	0.71	0.94	0.81	423
O	0.99	1.00	0.99	33973
accuracy			0.96	37687
macro avg	0.77	0.65	0.68	37687
weighted avg	0.97	0.96	0.96	37687

F1 score macro avg: 0.6788779838876352

Després d'haver probat d'afegir els 4 gazetteers un a un no hem millorat el millor f1 score avg que hem trobat anteriorment. Per tant, descartem l'opció d'afegir-los.

Context de la Paraula

Finalment, les últimes característiques que provarem d'afegir al nostre model són les característiques que hem establert per defecte de la paraula anterior i posterior de la paraula en la qual ens trobem. D'aquesta manera el CRF tindrà informació sobre les feature functions de les paraules que la rodegen. Si la paraula està a inici de frase, li posarem l'etiqueta 'BOS' i si es troba al final 'EOS'.

	precision	recall	f1-score	support
B-LOC	0.68	0.67	0.67	479
B-MISC	0.76	0.68	0.72	748
B-ORG	0.87	0.60	0.71	686
B-PER	0.64	0.74	0.69	703
I-LOC	0.51	0.30	0.38	64
I-MISC	0.51	0.52	0.51	215
I-ORG	0.83	0.68	0.75	396
I-PER	0.75	0.91	0.82	423
O	0.99	1.00	0.99	33973
accuracy			0.97	37687
macro avg	0.73	0.68	0.69	37687
weighted avg	0.97	0.97	0.97	37687

F1 score macro avg: 0.6939489870063816

Com es pot apreciar en els resultats, afegint les característiques de les paraules que rodegen la paraula a predir, millorem l'f1 score avg a 0.69395. Respecte al millor resultat que teníem fins ara hem aconseguit millorar notòriament B-MISC i I-MISC.

En resum, hem afegit les feature functions de POS-tags, lemmas, longitud de la paraula i context. Ara després d'haver ajustat el model anem a desfer la codificació BIO de les prediccions i anem a veure quin ha estat el millor resultat predint les entitats de les paraules sense la 'B' i la 'I'.

	precision	recall	f1-score	support
LOC	0.67	0.63	0.65	543
MISC	0.73	0.67	0.70	963
O	0.99	1.00	0.99	33973
ORG	0.87	0.64	0.74	1082
PER	0.71	0.83	0.76	1126
accuracy			0.97	37687
macro avg	0.79	0.75	0.77	37687
weighted avg	0.97	0.97	0.97	37687

F1 score macro avg: 0.7681105002217012

Anàlisis de resultats amb feature functions

Durant aquesta secció del treball, s'han dut a terme diverses modificacions i entrenaments del model Conditional Random Field (CRF) el qual hem utilitzat per entrenar un NER que fos capaç de reconèixer les entitats de Localització, Persona, Organització i Miscellaneous.

L'objectiu principal d'aquest apartat ha estat millorar els resultats de les prediccions del nostre model sobre la partició de validació (test_a) afegint i eliminant funcions característiques. L'anàlisi següent es basarà en com ha evolucionat el model fins a arribar als millors f1 scores macro avg que hem obtingut en els dos idiomes.

Les funcions característiques són regles o patrons que el CRF usa per capturar informació rellevant sobre les entitats que estem buscant identificar. En afegir noves funcions, s'esperava que el model pogués captar millor les relacions i els patrons del text, però ens hem trobat amb què afegir certes funcions que no són necessàries afegeixen soroll al CRF i porta a pitjors resultats. També cal comentar que l'ús de la mètrica f1 macro avg en ser una mesura no ponderada ens ha portat al fet que la millora en la predicció de totes les classes tingués el mateix valor, de manera que hem evitat un biaix en millorar únicament aquelles classes que apareixen més en el corpus, sinó que li hem donat el mateix valor que a aquelles que apareixen menys.

En resum, després d'analitzar els resultats, hem pogut veure com les modificacions que hem fet en les funcions característiques han tingut un impacte positiu en les prediccions del nostre model.

Codificacions

En aquest apartat, un cop s'ha fixat les característiques que aporten millors resultats al nostre model en una codificació donada, provarem diferents codificacions per a veure quina d'elles és més favorable. Nosaltres hem realitzat les proves sobre la codificació BIO i, posteriorment, hem provat el conjunt de característiques òptimes seleccionat sobre les codificacions IO i BIOW.

Com sabem, les codificacions tenen el següent significat i funcionament:

- BIO
    - B → *Beginning*, marca l'inici d'una entitat
    - I → *In*, marca que és dins d'una entitat, prèviament cal l'aparició d'una B
    - O → *Out*, es troba fora de qualsevol entitat
- IO
    - Presenta el mateix esquema que BIO, però en aquest cas no trobem B en la codificació, és a dir no marquem l'inici de les entitats.
- BIOW
    - Presenta el mateix esquema que BIOW, però en aquest cas, destaquem les paraules aïllades que són entitats amb una W. Per a que una paraula que representa una entitat es consideri aïllada, en la codificació BIO aquesta hauria d'estar etiquetada amb una B i estar seguida d'una paraula no que fos etiquetada amb I.

Anàlisis de resultats amb codificacions

Realitzada, doncs, aquesta explicació, compararem els resultats obtinguts amb les diferents codificacions usades, com hem mencionat anteriorment. Per a fer una comparació equitativa i justa, ens fixarem únicament en la entitat predita com a tal. És a dir, cadascun dels models tindrà més o menys classes segons la seva codificació, però nosaltres, per a comparar-los, ens centrarem únicament en si allò que han predit és LOC, MISC, ORG, PER o O. D'aquesta manera garantim que totes les codificacions s'avaluin sobre les mateixes classes i es puguin comparar equitativament.

Model Castellà

BIO

	precision	recall	f1-score	support
LOC	0.65	0.80	0.72	1321
MISC	0.52	0.54	0.53	1099
O	0.99	0.99	0.99	45356
ORG	0.82	0.73	0.77	3066
PER	0.84	0.83	0.84	2081
accuracy			0.96	52923
macro avg	0.76	0.78	0.77	52923
weighted avg	0.96	0.96	0.96	52923

F1 score macro avg: 0.7686763504778588

IO

	precision	recall	f1-score	support
LOC	0.62	0.77	0.69	1321
MISC	0.49	0.47	0.48	1099
O	0.99	0.99	0.99	45356
ORG	0.78	0.72	0.75	3066
PER	0.85	0.82	0.83	2081
accuracy			0.95	52923
macro avg	0.75	0.76	0.75	52923
weighted avg	0.95	0.95	0.95	52923

F1 score macro avg: 0.7487933568816227

BIOW

	precision	recall	f1-score	support
LOC	0.65	0.80	0.72	1321
MISC	0.50	0.55	0.53	1099
O	0.99	0.99	0.99	45356
ORG	0.84	0.73	0.78	3066
PER	0.85	0.84	0.84	2081
accuracy			0.96	52923
macro avg	0.77	0.78	0.77	52923
weighted avg	0.96	0.96	0.96	52923

F1 score macro avg: 0.7710175644671462

Com es pot apreciar, els millors resultats han estat obtinguts amb la codificació BIOW, on hem obtingut un f1 score de 0.771. Això pot ser degut a que la codificació BIOW presenta una major varietat de classes i, per tant, permet detectar millor les diferents entitats presents al text. També pot ser que el fet de diferenciar entre aquells termes que estan a inici de entitats composades de paraules i aquells que són entitats aïllades, ajudi al CRF a diferenciar alguna característica representativa de cada grup i que per tant classifiqui millor les entitats que altres codificacions que no diferencien entre aquests termes.

Concluim que la codificació BIOW és útil en casos on les entitats poden ser llargues o estar superposades dins les oracions. Com que l'espanyol es un idioma que presenta una elevada quantitat de noms propis i entitats amb les característiques mencionades anteriorment, la codificació BIOW és especialment útil en aquest cas.

Model Neerlandés

BIO

	precision	recall	f1-score	support
LOC	0.67	0.63	0.65	543
MISC	0.73	0.67	0.70	963
O	0.99	1.00	0.99	33973
ORG	0.87	0.64	0.74	1082
PER	0.71	0.83	0.76	1126
accuracy			0.97	37687
macro avg	0.79	0.75	0.77	37687
weighted avg	0.97	0.97	0.97	37687

F1 score macro avg: 0.7681105002217012

IO

	precision	recall	f1-score	support
LOC	0.65	0.62	0.63	543
MISC	0.73	0.64	0.68	963
O	0.99	1.00	0.99	33973
ORG	0.86	0.65	0.74	1082
PER	0.70	0.82	0.76	1126
accuracy			0.97	37687
macro avg	0.79	0.75	0.76	37687
weighted avg	0.97	0.97	0.97	37687

F1 score macro avg: 0.7616068138268444

BIOW

	precision	recall	f1-score	support
LOC	0.69	0.62	0.65	543
MISC	0.69	0.66	0.68	963
O	0.99	1.00	0.99	33973
ORG	0.87	0.62	0.72	1082
PER	0.70	0.83	0.76	1126
accuracy			0.97	37687
macro avg	0.79	0.75	0.76	37687
weighted avg	0.97	0.97	0.97	37687

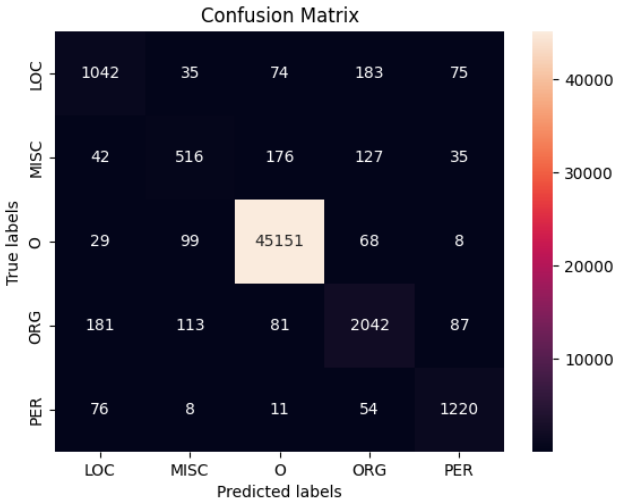
F1 score macro avg: 0.7606331089469076

Després de calcular els resultats amb els tres tipus de codificació sobre el corpus en neerlandés podem veure com amb la codificació BIO arribem als millors resultats. Això pot passar perquè la codificació BIO s'adequa de manera natural al model CRF, ja que la seva estructura seqüencial s'ajusta al concepte de dependències seqüencials que el model CRF

té en compte. Aquesta combinació millorar el rendiment del sistema NER perquè el model pot aprendre millor els patrons i capturar amb precisió els límits de les entitats en base al context seqüencial de les paraules. En el cas de la codificació BIOW pot ser que en aquest cas no hagi millorat perquè el model no ha trobat característiques representatives d'aquestes entitats aïllades i per tant el que ha fet ha estat afegir més soroll a les prediccions del model.

Predicció en partició de test_b amb millor codificació

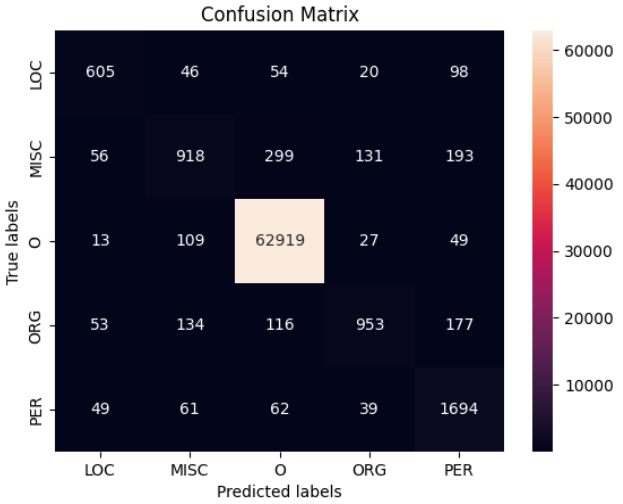
Model Castellà



	precision	recall	f1-score	support
LOC	0.76	0.74	0.75	1409
MISC	0.67	0.58	0.62	896
O	0.99	1.00	0.99	45355
ORG	0.83	0.82	0.82	2504
PER	0.86	0.89	0.87	1369
accuracy			0.97	51533
macro avg	0.82	0.80	0.81	51533
weighted avg	0.97	0.97	0.97	51533

F1 score macro avg: 0.8113371834302843

Model Holandés (BIO)



	precision	recall	f1-score	support
LOC	0.78	0.74	0.76	823
MISC	0.72	0.57	0.64	1597
O	0.99	1.00	0.99	63117
ORG	0.81	0.67	0.73	1433
PER	0.77	0.89	0.82	1905
accuracy			0.97	68875
macro avg	0.82	0.77	0.79	68875
weighted avg	0.97	0.97	0.97	68875

F1 score macro avg: 0.7894324289254558

Anàlisi de les prediccions de test

Si analitzem les prediccions fetes sobre la partició de test, podem veure com obtenim resultats similars als de la validació. Cal afegit que en les matrius de confusió es pot apreciar com la gran majoria de termes es prediuen com a entitats 'O', també es pot veure pel que fa les altres entitats, com els models les confonen força, sobretot la entitat MISC.

Etiquetatge de textos reals

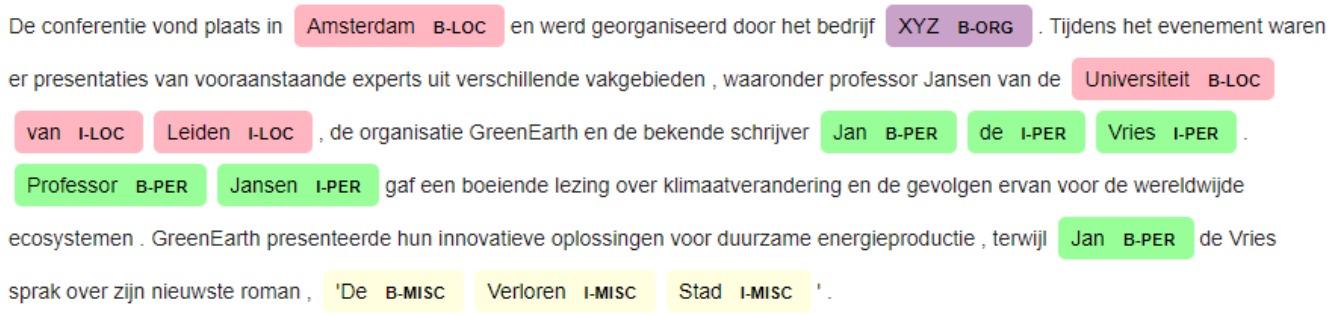
Per fer predicció de textos reals hem implementat la funció predict_raw_text , la qual rep com a paràmetre el text que volem etiquetar i retorna el mateix text amb les entitats reconegudes destacades amb colors. Per implementar aquesta funció hem fet ús de la llibreria spacy per etiquetar les paraules del text amb el seu postag i el seu lemma i de la funció displacy.render, que ens ha permès destacar les paraules que el nostre model ha etiquetat d'una manera visual.

Volem destacar que el format dels postags predits per spacy són diferents dels que venen donats al corpus d'entrenament tant en el castellà com en el neerlandès. En el cas de l'holandès la diferència està en el format de les lletres i hem pogut canviar-ho bé amb un diccionari. En canvi, en el cas del castellà veiem com el corpus d'entrenament té etiquetes molt més específiques de les que ens dona spacy sobre el text que volem predir, per tant, posarem un guió a la part del postag de manera que el model no tingui en compte el que posa. Si creéssim un diccionari com hem fet amb neerlandès estaríem afegint un postag fals que modificaria el comportament del model amb informació falsa, per exemple si construïm un diccionari que canvia de PUNCT a 'Fd' (signe de puntuació punt), estariem dient que aquell signe de puntuació és un punt i podria ser una coma ('Fc') o un guió ('Fg').

Text en Castellà

El famoso escritor peruano Mario B-PER Vargas I-PER Llosa I-PER asistió a la conferencia sobre literatura latinoamericana en la ciudad de Buenos B-LOC Aires I-LOC . Durante su discurso , destacó la importancia de la narrativa en la sociedad actual . Vargas B-PER Llosa I-PER es reconocido a nivel mundial por sus novelas , como 'La B-MISC ciudad I-MISC y los perros ' y 'Conversación B-MISC en I-MISC la I-MISC catedral ' . También ha sido galardonado con el Premio B-MISC Nobel I-MISC de I-MISC Literatura I-MISC en I-MISC 2010 I-MISC . El evento fue organizado por la Universidad B-ORG de I-ORG Buenos I-ORG Aires I-ORG y contó con la participación de diversos intelectuales y académicos . Entre los asistentes destacados se encontraban representantes de la editorial Alfaguara B-LOC y la Fundación B-ORG para I-ORG las I-ORG Letras I-ORG Mexicanas I-ORG .

Text en Neerlandés



En els textos etiquetats que tenim tant pel Castellà com pel Neerlandès podem apreciar com el nostre NER ha etiquetat algunes entitats correctament i altres incorrectament.

Pel que fa al text en castellà es pot veure com per exemple encerta en reconèixer Buenos Aires com una localització i Mario Vargas Llosa com una persona i falla quan prediu Alfaguara com una Localització quan és una organització i també quan es deixa per seleccionar 'y los perros' i 'catederal' com entitats MISC.

En el text en Neerlandès es poden veure encerts quan etiqueta Amsterdam com una localització o XYZ com una organització. Per altra banda, també es poden veure error, per exemple quan etiqueta a Jan de Vries sense el cognom de Vries o quan etiqueta amb ‘O’ GreenEarth quan es una organització.

Part Opcional

Introducció

Durant aquesta part del treball construirem un etiquetador d'entitats anomenades (NER) basat en Conditional Random Fields (CRF) de la mateixa manera que hem fet en l'anterior part del treball. La diferència és que aquest cop entrenarem el nostre CRF amb el corpus CADEC.

CSIRO Adverse Drug Event Corpus (Cadec) és un corpus etiquetat a partir de posts de fòrums mèdics sobre efectes adversos en els medicaments. El corpus està tret de publicacions de les xarxes socials i conté text en anglès que a vegades no segueix estrictament les normes gramaticals i sintàctiques. Cada paraula de cada document que ve en el corpus està etiquetada amb codificació BIO juntament amb les entitats nominals i un identificador basat en el coneixement científic del creador del corpus. Nosaltres per fer aquest treball no tindrem en compte l'identificador i treballarem únicament amb la codificació BIO i les entitats anomenades següents: Adverse Drug Reaction (ADR, efecte advers d'un medicament), Disease (Di, la raó per la qual es pren el medicament), Drug (Dr, nom del medicament), Symptom (S, manifestacions de la malaltia) i Finding (F, un terme clínic que no pertany a cap de les entitats anteriors).

Implementació

Per entrenar el nostre CRF el que hem fet ha sigut crear les funcions read_data i add_lemmas_postags les quals respectivament llegeixen els fitxers de train i test de les dades i els posen en el format adient perquè el nostre CRF pugui ser entrenat. Recordem que el nostre CRF està entrenat per rebre una llista de llistes on cada subllista representa un document i està formada per tuples de 4 elements on cada tupla té la forma (paraula, postag, lemma, BIO-label). Exemple del format: [[('pain', 'pain', 'NOUN', 'B-ADR'), ('in', 'in', 'ADP', 'I-ADR'),... . Cal destacar que per calcular els postags i els lemmas hem utilitzat la llibreria SpaCy.

Per ser rigorosos en l'entrenament del model la partició de train l'hem separat en val i train on val estarà formada per 150 documents i train per 768.

Feature Functions

En aquesta secció treballarem de la mateixa manera que en la primera part del treball. Afegirem i traurem feature functions de la funció words2features i anirem observant els resultats obtinguts sobre la partició de validació fins a quedar-nos amb aquelles que donin millors resultats.

Feature Functions	**F1 score avg macro**
Default	0.350723
Default + POS-tags	0.353101
Default + POS-tags + lemmas	0.358854
Default + POS-tags + lemmas + arrel	0.366615
Default + POS-tags + lemmas + arrel + longitud	0.362598
Default + POS-tags + lemmas + arrel + prefixos (fins llargada 3)	0.371219
Default + POS-tags + lemmas + arrel + prefixos (fins llargada 3) + features default del context	0.376790

Comparació default i amb millors features afegides:

	precision	recall	f1-score	support
O	0.94	0.97	0.95	25614
B-ADR	0.68	0.61	0.64	1427
I-ADR	0.61	0.57	0.59	2376
B-Di	0.33	0.19	0.24	83
I-Di	0.16	0.09	0.12	53
B-Dr	0.95	0.78	0.85	426
I-Dr	0.56	0.16	0.24	58
B-F	0.18	0.06	0.09	131
I-F	0.12	0.05	0.07	133
B-S	0.09	0.02	0.03	59
I-S	0.07	0.02	0.03	59
accuracy			0.90	30419
macro avg	0.43	0.32	0.35	30419
weighted avg	0.89	0.90	0.89	30419

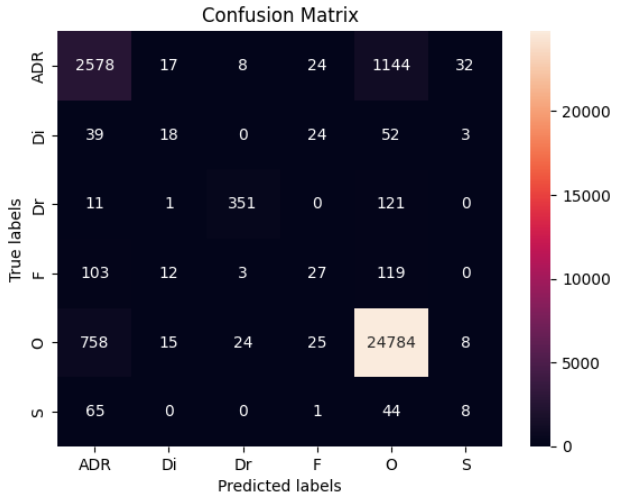
F1 score macro avg: 0.350723657655502

	precision	recall	f1-score	support
O	0.94	0.97	0.96	25614
B-ADR	0.68	0.64	0.66	1427
I-ADR	0.63	0.59	0.61	2376
B-Di	0.28	0.14	0.19	83
I-Di	0.25	0.09	0.14	53
B-Dr	0.93	0.79	0.85	426
I-Dr	0.48	0.19	0.27	58
B-F	0.30	0.12	0.17	131
I-F	0.21	0.08	0.11	133
B-S	0.22	0.10	0.14	59
I-S	0.08	0.03	0.05	59
accuracy			0.90	30419
macro avg	0.45	0.34	0.38	30419
weighted avg	0.89	0.90	0.90	30419

F1 score macro avg: 0.3767909090214798

En els dos informes de classificació que es poden veure tenim els resultats, en primer lloc, d'utilitzar les feature function que estan establertes per defecte en el nostre CRF i, en segon lloc, d'utilitzar les que ens han donat millors resultats. Es pot apreciar com hem millorat la F1 score de totes les categories excepte de B-Di, també es pot veure com els canvis més notables han succeït a les categories de B-F, I-F i B-S.

Si desfem la codificació BIO per avaluar com ha fet les prediccions el nostre model, utilitzant les feature functions que ens han anat millor, obtenim els següents resultats:

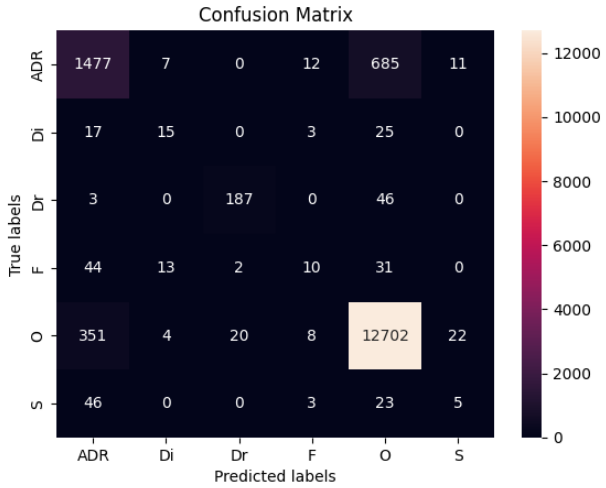


	precision	recall	f1-score	support
ADR	0.73	0.68	0.70	3803
Di	0.29	0.13	0.18	136
Dr	0.91	0.73	0.81	484
F	0.27	0.10	0.15	264
O	0.94	0.97	0.96	25614
S	0.16	0.07	0.09	118
accuracy			0.91	30419
macro avg	0.55	0.45	0.48	30419
weighted avg	0.90	0.91	0.91	30419

F1 score macro avg: 0.4811204054913585

Predicció en el Test

Podem veure com era d'esperar que en la predicció del test obtenim resultats molt similars als de la partició de validació.



	precision	recall	f1-score	support
ADR	0.76	0.67	0.72	2192
Di	0.38	0.25	0.30	60
Dr	0.89	0.79	0.84	236
F	0.28	0.10	0.15	100
O	0.94	0.97	0.95	13107
S	0.13	0.06	0.09	77
accuracy			0.91	15772
macro avg	0.57	0.48	0.51	15772
weighted avg	0.90	0.91	0.91	15772

F1 score macro avg: 0.5078508724967471

Anàlisi de resultats

Després d'haver provat i analitzat el nostre model, podem concloure que els resultats són poc fiables. En els diferents informes de classificació i les matrius de confusió hem pogut veure com aquelles categories que estan menys representades es prediuen amb un f1 score que sol estar clarament per sota del 0.5 % la qual cosa indica que estem cometent molts errors tant en la precisió com el recall.

D'aquesta anàlisi de resultats podem concloure que necessitem utilitzar alguna tècnica d'ampliació de corpus etiquetat per poder millorar els resultats del nostre model, ja que els errors són causats per la manca de mostres en entitats infrarepresentades.

Etiquetatge de textos reals

Sarah B-ADR experienced I-ADR acute I-ADR stomach I-ADR pain I-ADR after taking Diclofenac B-Dr for her chronic B-Di arthritis I-Di , which was the reason for taking the drug . The pain B-S was a clear adverse drug reaction . Additionally , she had been having trouble B-ADR sleeping I-ADR and frequently B-ADR felt I-ADR fatigued I-ADR , which were symptoms of her underlying disease . During her medical examination , the doctor noticed swelling B-ADR in I-ADR her I-ADR joints I-ADR , a finding that required further investigation . Moreover , Sarah 's blood tests revealed elevated B-F levels I-F of I-F liver I-F enzymes I-F , indicating a potential B-Di liver I-Di dysfunction I-Di , which was an unexpected finding .

En aquest text etiquetat pel nostre model CRF podem veure aquest acerta algunes prediccions de les quals fa, per exemple: Diclofenac com un drug, chronic arthritis com una disease o trouble sleeping com un ADR. Per altra banda, també comet errors quan per exemple identifica Sarah com un ADR.

Conclusions

En la primer part d'aquesta pràctica, hem desenvolupat dos sistemes d'extracció d'entitats anomenades (NER) utilitzant Conditional Random Fields (CRF) per al castellà i l'holandès. Hem fet servir el corpus conll2002 per entrenar i avaluar els nostres models. També hem explorat diverses característiques del text i provat les codificacions BIO, BIOW i IO en el corpus. Hem afegit funcions característiques com ara els Pos-Tags i els lemes per millorar les prediccions del model. Hem observat que l'ús de les arrels no han millorat els resultats i el context de la paraula només ha millorat en el Neerlandès. En conclusió, les característiques que han millorat la precisió i el recall dels nostres models CRF han estat: característiques per defecte, Pos-Tags, lemes, prefixos i context. Aquestes característiques ens han permès obtenir resultats més precisos en l'extracció d'entitats anomenades en textos en castellà i neerlandés. Respecte a les codificacions provades, per el model en neerlandés la que ha donat millor resultat ha estat BIO i en espanyol ha estat la codificació BIOW. Hem obtingut millors codificacions diferents en cadascun dels idiomes, això és degut a que cadascun d'ells presenta les seves característiques i estructura úniques. Es pot donar el cas, doncs on una codificació s'adapti millor a un idioma que a un altre a l'hora de fer les prediccions.

Pel que fa a la part opcional del treball, hem construït un etiquetador d'entitats anomenades (NER) utilitzant Conditional Random Fields (CRF) i el corpus CADEC. Hem millorat el rendiment del model afegint funcions característiques com el context, POS-tags i lemes d'entre altres. Tot i això, els resultats han estat poc fiables a causa de la falta de mostres per a les entitats menys representatives. Hem conclòs que cal fer ús tècniques d'ampliació del corpus etiquetat per millorar el model. En els textos reals, el CRF ha mostrat precisió en algunes prediccions, però també ha comès errors. És necessari continuar millorant el model per assolir resultats fiables.