

# Pràctica 3: Extracció d'entitats anomenades

Processament del Llenguatge Humà

**Joan Saurina i Ricós**

21783219B

**Sergi Tomàs Martínez**

39959144W



UNIVERSITAT POLITÈCNICA  
DE CATALUNYA  
BARCELONATECH

16 de maig de 2023

# Índex

<b>1</b>	<b>Introducció i objectius</b>	<b>2</b>
1.1	Introducció . . . . .	2
1.2	Objectius . . . . .	2
<b>2</b>	<b>Particionat de les dades</b>	<b>3</b>
<b>3</b>	<b>CRFTagger</b>	<b>3</b>
<b>4</b>	<b>Experimentació amb diferents codificacions</b>	<b>4</b>
<b>5</b>	<b>Funció d'avaluació</b>	<b>6</b>
<b>6</b>	<b>Avaluació</b>	<b>8</b>
6.1	Idioma espanyol . . . . .	8
6.1.1	Textos Reals: ESP . . . . .	10
6.2	Idioma neerlandès . . . . .	10
6.2.1	Textos Reals: NED . . . . .	12
<b>7</b>	<b>Conclusions Finals</b>	<b>13</b>
<b>8</b>	<b>Annex</b>	<b>14</b>
8.1	Models Idioma Espanyol . . . . .	14
8.2	Models Idioma Neerlandès . . . . .	15

# 1 Introducció i objectius

## 1.1 Introducció

L'extracció d'entitats anomenades és una tasca crucial en el camp de la lingüística computacional i la intel·ligència artificial, la qual s'enfoca a identificar i classificar noms propis del text, com ara noms de persones, organitzacions, ubicacions, expressions de temps o percentatges. La importància rau en la seva capacitat per entendre i estructurar informació no estructurada, facilitant tasques posteriors com ara la traducció automàtica, la resposta a preguntes, el resum de textos, entre d'altres.

En aquesta pràctica, el nostre objectiu és implementar un reconeixedor d'entitats anomenades utilitzant Conditional Random Fields (CRF), un tipus de model estadístic que es fa servir per predir una seqüència d'etiquetes per a una seqüència d'entrades. Farem ús de la classe `nltk.tag.CRFTagger` proporcionada per la biblioteca NLTK (Natural Language Toolkit).

## 1.2 Objectius

**Implementació de Reconixedors d'Entitats Anomenades:** S'ha desenvolupat dos reconixedors d'entitats nomenades, un per a l'idioma espanyol i un altre per al neerlandès, utilitzant CRFs.

**Experimentació amb funcions de característiques:** S'ha experimentat amb diferents funcions de característiques de `get_features` (funció de `CRFTagger`) per a provar de millorar el rendiment del model. El tractament de les *features* és essencial als CRFs, ja que ajuden a capturar diverses propietats del token, com ara la morfologia de les paraules, la informació del context, etc.

**Experimentació amb Diferents Codificacions:** Investigarem l'efecte de diferents codificacions d'etiquetes, incloent BIO (Beginning-Inside-Outside), BIOW (Beginning-Inside-Outside-Whole) i IO (Inside-Outside), en el rendiment del reconeixedor d'entitats nomenades.

**Avaluació dels Models:** Finalment, avaluarem el rendiment dels nostres models utilitzant un conjunt de prova i mesurarem la seva eficàcia mitjançant les mètriques pertinentment considerades, com ara la precisió, el recall i el F-Score per a models d'aquest tipus.

El resultat esperat d'aquesta pràctica és obtenir un reconeixedor d'entitats anomenades robust i eficaç per als idiomes espanyol i neerlandès, així com una experimentació en profunditat del paper que exerceixen les funcions de característiques i les diferents codificacions d'etiquetes als models de CRF.

## 2 Particionat de les dades

S'ha utilitzat la partició de Train, Validation i Test que ja proporciona la llibreria NLTK dels arxius de text *conll2022*.

S'ha transformat el format en que les dades venien implementades de manera que es troben separades la informació de la paraula de l'entitat a predir (Variable objectiu). Per a poder fer ús del CRFTagger i modificar les característiques evaluables, les tuples etiquetades del corpus tenen la següent estructura:  $((Word, POS_{Tag}), Entity)$ .

## 3 CRFTagger

S'ha construït una subclasse dins del *CRFTagger* de *NLTK* utilitzat per a la predicció, per a poder personalitzar la quantitat i el tipus d'informació que aquest usa. S'anomena CustomCRFTagger i concretament, s'ha reconstruït les funcions *get\_features()* i el mètode "privat" *\_get\_features()*. Es fa ús de les següents característiques:

- word: La paraula en sí.
- has\_punctuation: Booleana que comprova si el token conté un dels següents símbols: (",", "-", "?", "!")
- suffix: Les últimes 3 lletres o caràcters del token
- lenght: Longitud de la paraula.
- prefix: Les 3 primeres lletres o caràcters del token.
- lemma: Lema del token segons el lematitzador de *WordNet*
- is\_stopword: Booleana que comprova si el token es una *stop\_word*. Les stopwords en aquest cas són un set de paraules que proporciona NLTK que no aporten informació especialment rellevant per al POS\_tagging. Generalment, verbs i determinants.

A més a més, s'han parametritzat algunes features per a tal de poder experimentar el rendiment del model segons la seva implementació o no. Han estat:

- POS\_tag: Conté la informació POS\_tag del token ja inclosa al corpus.
- is\_uppercase: Booleana que indica si la paraula comença per lletra majúscula.
- pos\_pred: Indica el POS\_tag del token anterior
- pos\_post: Indica el POS\_tag del token següent.

## 4 Experimentació amb diferents codificacions

La codificació de les entitats és un aspecte clau del model, tant en la complexitat algorítmica que algunes propocionen com en la quantitat d'informació que aporten. S'ha experimentat amb diverses d'aquestes per a poder obtenir un millor enteniment del seu funcionament, propietats i resultats.

És important recalcar que les dades d'entrenament que s'han usat no contemplen entitats discontinúes i, per tant, el model tampoc serà capaç de predir-les. El fet d'afegir-ho implicaria donar-li al CRF una finestra de text al voltant del token molt gran i hagués augmentat considerablement els temps d'entrenament i predicció.

En primer lloc, la codificació BIO és un esquema utilitzat comunament en tasques de Reconeixement d'Entitats Nomenades (NER) i altres tasques d'etiquetatge de seqüències en Processament del Llenguatge Natural (NLP). Les dades d'entrenament proporcionades venien en aquest format, per el que el model base que s'ha entrenat treballa amb aquestes. BIO és un acrònim per a les tres categories d'etiquetes en aquest esquema:

- **B-** (Begin): Aquesta etiqueta es fa servir per marcar l'inici d'una entitat al text. Per exemple, si s'està intentant identificar persones en un text i el nom de la persona és "John Doe", "John" rebria l'etiqueta "B-PER".
- **I-** (Inside): Aquesta etiqueta es fa servir per marcar els tokens que estan dins una entitat, però no són el primer token. Seguint l'exemple anterior, Doe rebria l'etiqueta I-PER, indicant que és part de l'entitat PERSON que ha començat amb el token anterior.
- **O** (Outside): Aquesta etiqueta es fa servir per marcar els tokens que no formen part de cap entitat. Tots els tokens que no són persones, al nostre exemple, rebrien l'etiqueta "O".

Aquest esquema és útil perquè proporciona una manera d'identificar les entitats al text també quan les entitats poden constar de múltiples tokens. Quan s'utilitza les etiquetes "B-İ İ-", és possible delimitar exactament on comença i acaba cada entitat al text.

Per tal d'experimentar amb diferents codificacions s'ha implementat dues funcions: **convert\_to\_biow()** i **convert\_to\_io()**. El seu propòsit és transformar l'anotació d'entitats en un conjunt de dades de text en diferents formats de codificació. Aquests formats de codificació són utilitzats comunament en el camp d'Extracció d'Informació i Reconeixement d'Entitats Nomenades (NER), una àrea del Processament del Llenguatge Natural (NLP).

La primera funció, **convert\_to\_biow**, converteix les etiquetes al format BIO-W. Aquest format inclou quatre tipus d'etiquetes:

- **B-**: seguit del tipus important, que indica el començament d'una entitat.
- **I-**: seguit del tipus d'entitat, que indica que el token és part d'una entitat però no és el primer token de l'entitat.
- **W**: indica que el token no és part de cap entitat i el token anterior tampoc no ho era.

- **O**: indica que el token no és part de cap entitat però el token anterior sí que ho era.

La segona funció, **convert\_to\_io**, converteix les etiquetes al format IO. Aquest és un format més senzill amb només dos tipus d'etiquetes:

- **I-**: seguit del tipus important, que indica que el token és part d'una entitat. No distingeix entre l'inici d'una entitat i els tokens dins l'entitat.
- **O**: indica que el token no és part de cap entitat.

L'experimentació amb diferents codificacions té com a objectiu avaluar com les diferents maneres de representar la mateixa informació poden afectar el rendiment d'un sistema de NER. En alguns casos, un format de codificació pot ser més adequat que un altre per a un problema o conjunt de dades específic.

## 5 Funció d'avaluació

Es defineix una funció **compute\_scores** la qual s'encarrega de comparar les etiquetes reals d'un corpus i les predites pel model. A partir d'aquí retorna unes mètriques que descriuen com de bé funciona el model a l'hora d'identificar entitats.

En primer lloc, aquesta funció crida en crida a **get\_entities()**. Aquesta segona funció està dissenyada per extreure entitats d'un conjunt de dades basant-se en diferents esquemes de codificació, 'bio', 'biow' i 'io'. En termes generals, la funció itera sobre el conjunt de dades proporcionat (una llista de frases on cada frase és una llista de tuples, i cada tupla consta d'un element de la frase i la seva etiqueta corresponent). Després, s'examinen les etiquetes per determinar l'inici i el final de cada entitat, segons l'esquema de codificació donat. Les entitats identificades s'agreguen a una llista en forma d'una tupla, que conté: l'índex de la frase al conjunt de dades (representant el document), l'índex d'inici de l'entitat, l'índex final de l'entitat i la classe de l'entitat. Aquesta llista d'entitats es torna al final. Com ja s'ha mencionat anteriorment, aquest sistema de codificació no contempla entitats discontinües, per el que necessitaria canvis en cas de voler ser utilitzada en casos de NER més complexes.

Una vegada s'han coleccionat les entitats predites pel model, la funció passa a evaluar la qualitat d'aquestes. Al no tractar-se d'una predicció corrent, sinó que entren en jocs termes intermitjos (com identificar una entitat a mitjes o trobar-la però classificar-la incorrectament), s'han definit una sèrie de mètriques per a poder evaluar els resultats.

A continuació, es presenta una breu descripció d'aquestes [1]:

- **Entitats correctes** (CA - *correct assignment*): es comptabilitzen quan l'entitat del fitxer original coincideix exactament amb una entitat predita en els valors d'inici i finalització, així com en el tipus d'entitat i la frase en la que es troba.
- **Entitats incorrectes** (IA - *incorrect assignment*): es comptabilitzen quan els valors d'inici i final coincideixen, però no el tipus.
- **Entitats parcials** (PA - *partial assignment*): es comptabilitzen quan hi ha una intersecció no buida entre dos intervals [inici, final]. És a dir, el model ha trobat la entitat, la ha classificat correctament però no en el seu rang complet. En cas de detectar 2 intervals separats que pertanyen a la mateixa entitat real, el model només en contabilitza 1.
- **Entitats que falten** (MA - *missing assignment*): es comptabilitzen les entitats que apareixen al fitxer original però no a la predicció.
- **Entitats espúries** (SA - *spurious assignment*) es comptabilitzen quan l'entitat apareix al fitxer de predicció però no a l'original. És a dir, quan el model ha detectat una entitat que no existeix.

A partir d'aquestes definicions, calculem la *precisió*, el *recall* i el *F-score* estàndard de la següent manera:

$$\text{Rec} = \frac{CA + \frac{1}{2}PA}{CA + IA + PA + MA}$$

$$\text{Prec} = \frac{CA + \frac{1}{2}PA}{CA + IA + PA + SA}$$

$$\text{F1} = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$$



## 6 Avalució

En ambdós idiomes, s'apliquen primerament les conversions i es crea un corpus d'entrenament, validació i test per cadascuna de les tres codificacions explicades anteriorment.

Seguidament es fa un bucle on es prova cada codificació amb combinacions dels diferents paràmetres sintonitzables comentats a l'**apartat 3** de la funció CustomCRFTagger, que resulten en un total de 48 possibles models diferents.

En els següents apartats d'avaluen els models de cada idioma segons f-score, i la complexitat computacional del model, que serà determinada mitjançant el temps d'entrenament i de predicció (validació).

### 6.1 Idioma espanyol

El millor model segons els resultats és el que utilitza la codificació 'bio', amb l'etiquetatge POS activat, distinció en majúscules i etiquetes tant pre com post POS en compte. Aquesta configuració específica dona com a resultat la puntuació F alta amb temps raonables d'entrenament i validació:

Cod	Recall	Prec.	F-score	Pos	Upper	Pos_Pred	Pos_Post	Ent.	Val.
bio	0.749	0.721	0.735	T	T	T	T	141.530	13.288

Taula 1: Model idioma espanyol amb major rendiment

Diversos factors poden contribuir a aquests resultats:

**Etiquetatge POS:** habilitar l'etiquetatge POS pot fer una diferència significativa en el rendiment perquè les etiquetes POS sovint poden proporcionar un context important per entendre el paper d'una paraula en una frase. Moltes paraules tenen significats diferents segons el seu ús en una frase, i el seu POS pot ajudar el model a desambiguar aquests casos. Per exemple, la paraula "córrer" podria ser un verb o un substantiu, i el significat pot ser força diferent en cada cas.

**Distinció de majúscules:** el model ha mostrat un millor rendiment quan diferencia entre majúscules i minúscules. Això podria ser degut al fet que el casing pot proporcionar informació important en molts idiomes, inclòs el castellà. Per exemple, els noms, els llocs i l'inici de les frases solen estar en majúscula. Aquest context addicional podria ajudar el model a distingir entre diferents usos de la mateixa paraula.

**POS Pred i POS Post:** la inclusió de les etiquetes POS anteriors i següents com a funcions sembla ser beneficiosa per al model. Això té sentit, ja que el llenguatge és inherentment seqüencial i el significat d'una paraula sovint pot dependre de les paraules que l'envolten. Tenint en compte les etiquetes POS de paraules adjacents, el model pot capturar millor l'estructura sintàctica de la frase, la qual cosa pot conduir a prediccions més precises.

**Codificació BIO:** aquesta estratègia d'etiquetatge (que significa Beginning, Inside i Outside) ajuda el model a comprendre no només el tipus d'entitat anomenada, sinó

també on comença i acaba. Això podria ser especialment útil per reconèixer entitats de diverses paraules, millorant així el rendiment general.

L'equilibri entre la puntuació F, el temps d'entrenament i el temps de validació suggereix que la configuració d'aquest model no només és precisa sinó també eficient. És capaç d'aprendre eficaçment de les dades d'entrenament i generalitzar ràpidament el seu aprenentatge a dades de validació no vistes.

Visualment, els diferents models es veuen de la manera següent:

Codificacions de models (Espanyol)

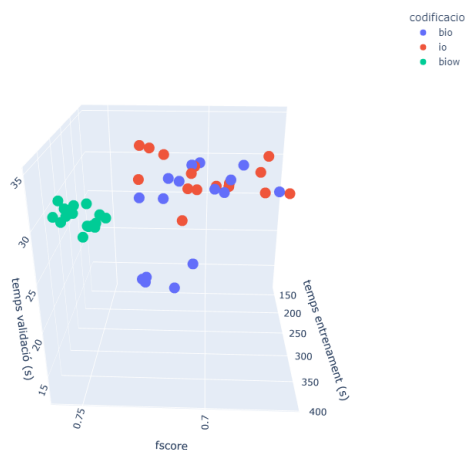


Figura 1: Plot en 3 dimensions dels models entrenats respecte temps d'entrenament, validació i F-Score

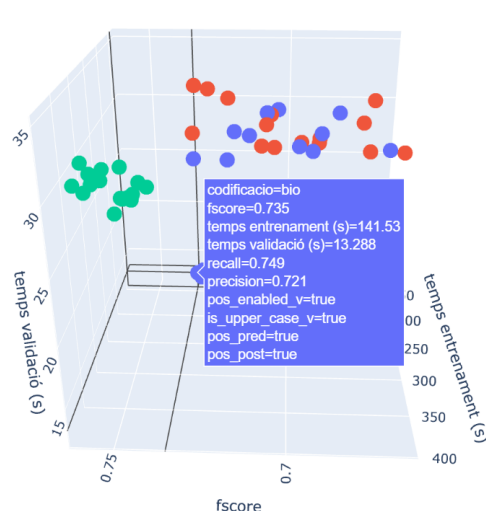


Figura 2: Posició i característiques del model escollit al gràfic

Pel que fa als clusters observats al diagrama de dispersió 3D, cal destacar que la codificació *BIOW* forma majoritàriament un clúster diferent representat en verd, mentre que les codificacions "bio i io" es barregen, representades per blau i vermell, respectivament.

La codificació *BIOW* podria estar formant un clúster separat a causa de la manera diferent en què gestiona les etiquetes inicials i interiors d'entitats de diverses paraules en comparació amb *BIO* i *IO*, donant lloc a diferents característiques de rendiment del model. D'altra banda, les codificacions "bio i io" podrien estar entrelaçades a causa dels seus esquemes d'etiquetatge similars, donant lloc a rendiments similars del model i, per tant, la seva presència superposada al diagrama de dispersió.

### Test:

Un cop s'ha escollit el model que ofereix el millor rendiment, es prova a les dades de test:

Cod	Recall	Prec.	F-score	Pos	Upper	Pos_Pred	Pos_Post	Ent.	Val.
bio	0.798	0.782	0.790	T	T	T	T	315.33	30.92

Taula 2: Model d'idioma espanyol amb major rendiment a les dades de test

Els resultats mostren força homogeneïtat en quant a resultats de rendiment entre les codificacions *BI* i *BIO*. Tot i que els models de la codificació *BIOW* mostren certa milloria en quant a resultats, aquesta no és gaire significativa i menys si es té en compte la diferència substancial en quant a temps d'entrenament.

Els resultats complets de tota la experimentació es troben tant al notebook adjunt amb l'entrega d'aquest document com a l'Annex, a l'apartat **8.1**.

### 6.1.1 Textos Reals: ESP

Per tal de comprovar la capacitat de generalització del model, se l'ha testejat amb una sèrie d'oracions que surten del àmbit directe amb el que ha estat entrenat, com és de les notícies. Cal ressaltar que aquest àmbit és extens i molt divers, complet i tracta amb molts tòpics diferents.

Les oracions han estat creades mitjançant un *LLM* com és *GPT-4*, demanant-li que les generi amb el tipus d'entitats que el model ha vist. S'ha valorat utilitzar una base de dades més extensa per a realitzar una experimentació més completa, però no se n'ha trobat cap que contingues únicament les entitats amb les que el model ha estat entrenat. L'ús d'un *LLM* ha estat de molta utilitat. Les oracions que s'han utilitzat per a l'experimentació es troben al notebook a l'apartat corresponent.

Els resultats han estat els següents:

Recall	Prec.	F-score
0.78	0.78	0.78

Taula 3: Resultats obtinguts per al model espanyol.

Els resultats indiquen que el model és capaç de funcionar correctament amb textos diferents als de la base de dades d'entrenament i predicció principals.

## 6.2 Idioma neerlandès

El model òptim per a l'idioma holandès, basat en la F-score, el temps d'entrenament i el temps de validació, es configura de manera diferent al model espanyol. El millor model holandès utilitza la codificació "biow", i els paràmetres de funció que POS habilitat s'estableixen com a fals, majúscules en vertader, pos\_pred en vertader i pos\_post en fals:

Cod	Recall	Prec.	F-score	Pos	Upper	Pos_Pred	Pos_Post	Ent.	Val.
biow	0.760	0.733	0.746	F	T	T	F	168.064	13.429

Taula 4: Model idioma neerlandès amb major rendiment

Aquesta diferència en la configuració òptima entre idiomes subratlla el fet que les característiques específiques de l'idioma poden influir en l'eficàcia de diferents característiques en un model de reconeixement d'entitats amb nom (NER).

Quan POS habilitat s'estableix en fals, indica que l'etiquetatge de part de veu (POS) no és beneficiós per a la tasca en el context de l'idioma holandès. Això podria ser degut a

les característiques específiques de l'idioma de l'holandès, on la utilitat de la informació del TPV no és tan pronunciada com en espanyol per a la tasca de NER.

La característica majúscula es manté com a vertadera, com en el model espanyol òptim, fet que suggereix que la capitalització d'entitats nomenades és una característica probablement universal beneficiosa per a les tasques NER en diferents idiomes.

El paràmetre `pos_pred` que s'estableix en *True* indica que el model es beneficia de tenir en compte l'etiqueta POS de la paraula anterior. D'altra banda, el paràmetre `pos_post` que s'estableix a *false* suggereix que tenir en compte l'etiqueta POS de la paraula següent no contribueix significativament al rendiment del model per a l'idioma holandès.

Finalment, la codificació *BIOW* que és la millor per a l'holandès, mentre que "bio" era òptima per a l'espanyol, suggereix que la manera com es gestionen les entitats de diverses paraules (*BIOW* utilitza una etiqueta "W", independent per a entitats de diverses paraules) ser més adequat per a la llengua holandesa, a causa de les seves característiques lingüístiques o de les especificitats del conjunt de dades utilitzat.

Visualment, els diferents models es veuen de la manera següent:

Codificacions de models (Neerlandès)

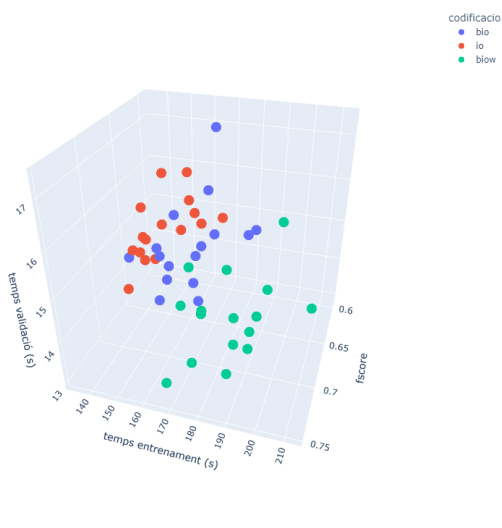


Figura 3: Plot en 3 dimensions dels models entrenats respecte temps d'entrenament, validació i F-Score

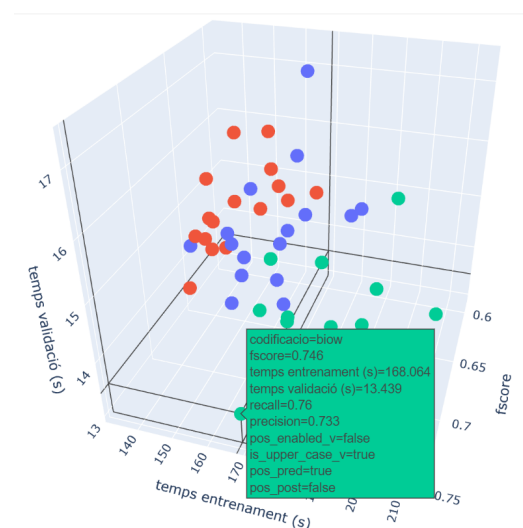


Figura 4: Posició i característiques del model escollit al gràfic

Igual que amb el model espanyol, s'observen formacions de clúster, amb "biow" formant un clúster diferent i les codificacions "bio" i "io" entremesclades.

### Test:

Un cop s'ha escollit el model que ofereix el millor rendiment, es prova a les dades de test:

Els resultats, novament, mostren força homogeneïtat en quant a resultats de rendiment entre les codificacions *BI* i *BIO*. Tot i que els models de la codificació *BIOW* mostren

Cod	Recall	Prec.	F-score	Pos	Upper	Pos_Pred	Pos_Post	Ent.	Val.
bio	0.771	0.761	0.768	F	T	T	F	222.57	42.26

Taula 5: Model d'idioma neerlandès amb major rendiment a les dades de test

certa milloria en quant a resultats, aquesta no és gaire significativa i menys si es té en compte la diferència substancial en quant a temps d'entrenament.

Els resultats complets de tota la experimentació es troben tant al notebook adjunt amb l'entrega d'aquest document com a l'Annex, a l'apartat **8.2**.

### 6.2.1 Textos Reals: NED

Els textos han estat generats novament amb CHAT-GPT.

Els resultats han estat els següents:

Recall	Prec.	F-score
1.00	0.80	0.88

Taula 6: Resultats obtinguts per al model neerlandès

Els resultats indiquen que el model és capaç de funcionar correctament amb textos diferents als de la base de dades d'entrenament i predicció principals.

## 7 Conclusions Finals

L'experimentació que s'ha realitzat ens ha permès comprendre en més profunditat el funcionament els CRFs, i més concretament, en la tasca del *Name Entity Recognition*.

El més destacable d'aquest procés ha estat veure la relació directe entre la complexitat i quantitat d'informació amb la que el model treballava, i el seu rendiment. També s'ha vist que, tot i que més complexitat otorgui millors resultats, en casos com el de la codificació *BIOES*, la complexitat computacional d'aquests pot arribar a ser contraproductiva, ja que incrementa en gran mesura els temps d'entrenament a canvi de poca millora en els resultats.

Depenent de la importància de la tasca a realitzar pot ser millor assumir els costos que aquest tipus d'implementació comporten, o mantenir resultats lleugerament pitjors a canvi de temps de computació més raonables.

S'han pogut apreciar també diferències substancials en el rendiment dels models al canviar d'idioma, demostrant que arribar a una configuració universal d'aquest tipus de models no és viable i s'ha de ser extremadament curós i tenir en compte el context de la tasca a realitzar.

## 8 Annex

### 8.1 Models Idioma Espanyol

<b>cod</b>	<b>recall</b>	<b>precision</b>	<b>fscore</b>	<b>pos_enabled</b>	<b>upper</b>	<b>pred</b>	<b>post</b>	<b>ent.</b>	<b>val.</b>
bio	0.749	0.721	0.735	True	True	True	True	141.530	13.288
bio	0.774	0.667	0.717	True	False	True	True	148.793	12.437
bio	0.748	0.718	0.733	False	True	True	True	134.963	13.280
bio	0.772	0.652	0.707	False	False	True	True	143.593	15.777
bio	0.750	0.717	0.733	True	True	True	False	149.333	13.236
bio	0.776	0.644	0.704	True	False	True	False	205.594	30.957
bio	0.748	0.709	0.728	False	True	True	False	309.711	30.235
bio	0.776	0.627	0.694	False	False	True	False	316.779	31.121
bio	0.733	0.703	0.718	True	True	False	True	320.772	30.561
bio	0.766	0.641	0.698	True	False	False	True	336.906	32.070
bio	0.728	0.696	0.712	False	True	False	True	303.854	31.774
bio	0.763	0.624	0.687	False	False	False	True	318.155	33.824
bio	0.733	0.700	0.716	True	True	False	False	318.083	32.462
bio	0.777	0.624	0.692	True	False	False	False	334.937	32.904
bio	0.725	0.690	0.707	False	True	False	False	304.179	33.395
bio	0.778	0.592	0.673	False	False	False	False	325.997	31.561
io	0.744	0.714	0.728	True	True	True	True	293.387	34.937
io	0.760	0.661	0.707	True	False	True	True	325.620	33.182
io	0.744	0.715	0.729	False	True	True	True	288.300	31.425
io	0.760	0.644	0.697	False	False	True	True	291.231	30.948
io	0.740	0.708	0.724	True	True	True	False	295.761	34.769
io	0.759	0.636	0.692	True	False	True	False	307.039	31.746
io	0.739	0.701	0.719	False	True	True	False	267.339	33.417
io	0.755	0.613	0.677	False	False	True	False	303.067	34.285
io	0.727	0.696	0.711	True	True	False	True	284.385	27.017
io	0.754	0.640	0.692	True	False	False	True	298.374	31.154
io	0.721	0.692	0.706	False	True	False	True	274.859	32.449
io	0.748	0.618	0.676	False	False	False	True	270.818	29.672
io	0.723	0.687	0.705	True	True	False	False	260.007	29.565
io	0.755	0.615	0.678	True	False	False	False	259.424	31.494
io	0.730	0.688	0.709	False	True	False	False	253.736	29.460
io	0.760	0.591	0.665	False	False	False	False	259.938	29.282
biow	0.763	0.750	0.756	True	True	True	True	358.554	29.398
biow	0.780	0.713	0.745	True	False	True	True	393.254	29.382
biow	0.759	0.744	0.752	False	True	True	True	342.113	30.465
biow	0.779	0.714	0.745	False	False	True	True	370.355	29.546
biow	0.760	0.745	0.752	True	True	True	False	349.191	29.939
biow	0.776	0.710	0.742	True	False	True	False	372.911	29.531
biow	0.756	0.741	0.748	False	True	True	False	329.315	30.220
biow	0.777	0.713	0.744	False	False	True	False	360.137	29.190
biow	0.769	0.752	0.760	True	True	False	True	347.700	29.478

biow	0.781	0.712	0.745	True	False	False	True	363.959	29.242
biow	0.767	0.748	0.757	False	True	False	True	329.377	29.687
biow	0.779	0.707	0.742	False	False	False	True	343.725	29.667
biow	0.765	0.748	0.756	True	True	False	False	335.816	29.158
biow	0.776	0.711	0.743	True	False	False	False	353.186	29.180
biow	0.771	0.749	0.760	False	True	False	False	321.317	30.190
biow	0.777	0.707	0.740	False	False	False	False	339.634	29.197

## 8.2 Models Idioma Neerlandès

<b>cod</b>	<b>recall</b>	<b>precision</b>	<b>fscore</b>	<b>pos_enabled</b>	<b>upper</b>	<b>pred</b>	<b>post</b>	<b>ent.</b>	<b>val.</b>
bio	0.721	0.675	0.697	True	True	True	True	175.397	16.827
bio	0.809	0.535	0.644	True	False	True	True	182.619	15.206
bio	0.719	0.663	0.690	False	True	True	True	161.811	16.131
bio	0.833	0.502	0.627	False	False	True	True	183.599	15.083
bio	0.724	0.673	0.698	True	True	True	False	159.885	14.807
bio	0.815	0.529	0.641	True	False	True	False	169.198	17.415
bio	0.713	0.657	0.684	False	True	True	False	153.409	13.967
bio	0.835	0.509	0.632	False	False	True	False	159.814	13.019
bio	0.689	0.642	0.665	True	True	False	True	153.850	14.496
bio	0.809	0.504	0.621	True	False	False	True	165.433	14.720
bio	0.692	0.634	0.662	False	True	False	True	149.510	14.651
bio	0.842	0.476	0.609	False	False	False	True	154.866	13.847
bio	0.684	0.632	0.657	True	True	False	False	147.281	14.748
bio	0.807	0.498	0.616	True	False	False	False	154.812	13.192
bio	0.680	0.621	0.649	False	True	False	False	132.937	14.237
bio	0.837	0.482	0.611	False	False	False	False	157.902	14.180
io	0.706	0.669	0.687	True	True	True	True	149.257	14.965
io	0.779	0.521	0.624	True	False	True	True	157.509	15.232
io	0.706	0.657	0.681	False	True	True	True	147.536	16.048
io	0.795	0.492	0.608	False	False	True	True	151.643	15.981
io	0.691	0.648	0.669	True	True	True	False	136.702	13.786
io	0.777	0.518	0.621	True	False	True	False	138.396	13.794
io	0.693	0.642	0.667	False	True	True	False	145.323	15.099
io	0.794	0.489	0.605	False	False	True	False	147.705	14.437
io	0.679	0.639	0.658	True	True	False	True	137.157	14.610
io	0.793	0.509	0.620	True	False	False	True	168.788	15.153
io	0.675	0.622	0.648	False	True	False	True	137.828	14.406
io	0.796	0.472	0.593	False	False	False	True	155.037	14.500
io	0.678	0.637	0.657	True	True	False	False	150.073	15.352
io	0.784	0.503	0.613	True	False	False	False	153.183	15.367
io	0.671	0.622	0.646	False	True	False	False	139.117	14.782
io	0.795	0.469	0.590	False	False	False	False	136.557	15.646
biow	0.761	0.737	0.749	True	True	True	True	196.151	14.725
biow	0.768	0.699	0.732	True	False	True	True	213.272	15.434
biow	0.753	0.728	0.741	False	True	True	True	180.061	15.158



biow	0.783	0.695	0.736	False	False	True	True	203.455	17.001
biow	0.758	0.732	0.745	True	True	True	False	190.988	15.224
biow	0.761	0.700	0.729	True	False	True	False	199.327	15.591
biow	0.760	0.733	0.746	False	True	True	False	168.064	13.439
biow	0.765	0.684	0.722	False	False	True	False	185.373	15.716
biow	0.745	0.718	0.731	True	True	False	True	174.578	13.671
biow	0.766	0.686	0.724	True	False	False	True	185.941	13.390
biow	0.752	0.723	0.737	False	True	False	True	179.235	14.999
biow	0.812	0.681	0.741	False	False	False	True	190.498	14.560
biow	0.750	0.719	0.734	True	True	False	False	174.952	15.864
biow	0.758	0.682	0.718	True	False	False	False	195.126	14.757
biow	0.733	0.707	0.720	False	True	False	False	169.024	14.721
biow	0.797	0.679	0.733	False	False	False	False	194.659	

## Referències

- [1] **eHealthKD**. (2021). Tasks. Recuperat el 15 de maig de 2023, de <https://ehealthkd.github.io/2021/tasks>