Facultat d'Informàtica de Barcelona

GIA UPC

Processament del llenguatge humà

Pràctica 3

Alumnes :
Casanovas Poirier, Anna
Pumares Benaiges, Irene

Tutors :
Turmo Borrás, Jordi
Medina Herrera, Salvador

May 14, 2024

Contents

1 Introducció

1.1 Motivació del treball

La identificació d'entitats anomenades és molt important per a diverses aplicacions en el camp del processament del llenguatge natural. Permet identificar i classificar entitats significatives, com persones, organitzacions, llocs i altres elements en un text. Aquest anàlisis és molt útil per a tasques com la traducció automàtica, l'anàlisis de sentiments, la recerca d'informació en un text...

Els Conditional Random Fields són una opció molt potent per a poder capturar les relacions i dependències entre les paraules d'un text. A més a més, permeten considerar l'ús de múltiples funcions de característiques per millorar la precisió dels resultats, el que el fa una eina molt útil per aquesta tasca, la qual ens ofereix l'oportunitat d'aprendre sobre aquest algorisme i poder-lo aplicar a un problema real.

1.2 Descripció de la tasca i les dades

L'objectiu d'aquesta pràctica és desenvolupar un identificador d'entitats anomenades utilitzant Conditional Random Fields (CRF) per a l'idioma espanyol i neerlandès.

Per a l'entrenament i l'avaluació, es faran servir el conjunt de dades del corpus 'conll2022', el qual conté textos en els dos idiomes etiquetats amb entitats anomenades en format BIO. Aquest corpus inclou textos dividits en conjunts d'entrenament, validació i prova.

Amb aquestes dades, s'han d'entrenar els models CRF per cada llengua, fent ús de diverses codificacions i funcions característiques i avaluar-ne el rendiment utilitzant mètriques com la precisió i F1-score.

2 Part obligatòria

2.1 Implementació del codi

Per a implementar aquesta primera part, s'ha creat una classe anomenada CRFModel, amb diversos mètodes que podran ser cridats a posteriori. Els mètodes inclouen l'entrenament del model, la predicció d'etiquetes i l'avaluació del rendiment. A més proporcionen funcions per a treballar amb diferents codificacions. La classe també permet especificar una funció de característiques personalitzada per a afinar el rendiment del model.

Inicialització

La classe té cinc paràmetres:

- `train`: dades del train
- `test`: dades del test (o val)
- `model_file`: cadena de text que especifica el nom del fitxer on es guardarà el model CRF entrenat (ruta)
- `features`: funció de característiques personalitzada per al CRF
- `encoding`: esquema de codificació per a les etiquetes. Pot ser 'bio', 'bioes' o 'io'

Transformació de les dades

El mètode `transformar_dades` converteix les etiquetes segons l'esquema de codificació especificat (utilitzant la funció `to_bioes` o `to_io`, segons sigui necessari).

Mètodes d'entrenament i predicció

El mètode `train` permet entrenar el model CRF amb un conjunt de dades, després d'haver-les convertit a la codificació triada.

El mètode `predict` pren un conjunt de frases i utilitza el model entrenat per predit les etiquetes corresponents. Si es proporciona el nom del fitxer durant l'inicialització pot carregar el model des del fitxer.

Codificació i Extracció d'Entitats

La classe inclou dos mètodes per convertir etiquetes entre esquemes diferents. El mètode `to_bioes` converteix etiquetes BIO a BIOES, mentre que el mètode `to_io` converteix etiquetes BIO a IO. Són mètodes essencials per a adaptar el model a diferents esquemes de codificació.

El mètode `extraccio_entitats` s'encarrega d'extreure entitats de frases etiquetades. Utilitza l'esquema de codifiació per determinar com s'han d'interpretar les etiquetes per

identificar les entitats. Depenent de la codifiació utilitzada, el mètode s'adapta per extreure les entitats corresponents. El mètode retorna una tupla, on ens diu el índex de la posició on comença i acaba l'entitat i també ens dona el tipus d'entitat.

Avaluació i Matriu de Confusió

Per avaluar el rendiment del model, la classe té el mètode `evaluacio_entitats`, que calcula mètriques com recall, precisió i F1-Score basant-se en les entitats reals i les prediccions. Per fer-ho, compara els dos conjunts per calcular True Positives (TP), False Positives (FP), i False Negatives (FN).

A més, el mètode `matriu_confusio` crea una matriu de confusió que mostra les comparacions entre les entitats predites i les reals. Permet visualitzar la matriu i proporciona una representació clara dels errors de classificació.

Funció de característiques

La classe inclou un paràmetre anomenat `features`, que és una funció utilitzada per crear les característiques per a cada token en que aporten informació. La funció creada, `creacio_feature_function`, inclou les següents característiques:

- `lemma_pos_tags`: S'utilitzen etiquetes POS per a cada paraula en el text, convertintles a format WordNet per a la lematització. Aporten informació sobre el lema de la paraula i la seva etiqueta POS.
- `word_from`: Informació sobre la forma de la paraula, així com també la paraula en minúscules.
- `prefix_suffix`: Informació sobre els prefixos i sufixos de les paraules, cosa que pot ser útil per detectar patrons morfològics.
- `morphology`: Informació sobre la morfologia de les paraules, com si comencen amb majúscula, si són tot en majúscules, si són tot en minúscules, si contenen guions, si són numèriques, entre altres.
- `length`: Infromació sobre la longitud de les paraules.
- `position`: Informació sobre si el token és el primer o l'últim de la frase, ja que això pot ser rellevant per al context.
- `context`: Informació del token anterior i del següent, incloent la forma de la paraula, si comença amb majúscula, si està tot en majúscules o minúscules, entre altres.

La funció `features_personalitzada` serveix per especificar quines característiques es volen usar en el model. Conté un diccionari, on cada clau correspon a una possible característica i el valor associat indica si es vol utilitzar o no (True o False). Aquesta mateixa funció és l'encarregada de cridar a `creacio_feature_function`, que utilitza les característiques especificades.

Funció d'entrenament

La funció `entrenament_model` entrena i avalua el model amb diferents opcions d'encodings i funcions de característiques, és a dir, s'encarrega de cridar a totes les funcions i mètodes de la classe CRFModel.

2.2 Experiments

En aquest apartat, es realitzaran diversos experiments per trobar la millor combinació de paràmetres per obtenir el model final. En aquests experiments, s'utilitzarà la partició de validació per ajustar els paràmetres del model. Això ens permetrà provar el model final amb la partició de prova per avaluar-ne el rendiment.

Experiment 1. Sense features

La taula mostra els resultats de les mètriques obtingudes per a l'espanyol i el neerlandès, amb tres codificacions diferents (BIO, IO i BIOES). En aquest cas, s'ha executat el model sense funcions de característiques addicionals, és a dir, utilitzant només les característiques predeterminades. A continuació s'analitzen els resultats obtinguts:

	Idioma	Recall	Precisió	F1-Score
BIO	Espanyol	0.6621	0.7086	0.6846
	Neerlandès	0.5833	0.6483	0.6141
BIOES	Espanyol	0.6767	0.6999	0.6881
	Neerlandès	0.6109	0.6692	0.6387
IO	Espanyol	0.6344	0.6873	0.6598
	Neerlandès	0.5562	0.6497	0.5993

Table 1: Mètriques per al model sense feature_functions

En primer lloc, l'espanyol té una tendència general a tenir mètriques lleugerament més altres que el Neerlandès en totes tres codificacions. Aixo indica que el model s'adapata millor a les característiques d'aquest idioma.

Per altra banda, en termes de codificacions, BIOES sembla proporcionar els millors resultats, tant per a l'espanyol com per al neerlandès. Aquest fet ens suggereix que aquesta codificació és més efectiva. BIO presenta mètriques una mica més baixes i IO sembla ser el més febles de tots, probablement per la seva incapacitat per identificar els inicis i finals de les entitats, probablement perquè no és capaç d'identificar els inicis i finals de les entitats.

Per als següents experiment s'usarà només la codificació BIO i BIOES, ja que són les dues que han proporcionat millors resultats.

Experiment 2. Amb features

1 Només amb la característica `context`

	Idioma	Recall	Precision	F1-Score
BIO	Espanyol	0.2983	0.5990	0.3983
	Neerlandès	0.1495	0.6015	0.2394
BIOES	Espanyol	0.3934	0.6628	0.4937
	Neerlandès	0.2015	0.6222	0.3044

Table 2: Mètriques per al model amb `context`

En aquest cas, els dos idiomes són millors amb la codificació BIOES. Tot i així, s'observa una reducció respecte l'experiment anterior, on no s'utilitzava cap funció de característiques addicionals. Podem concloure que amb la informació del token anterior i següent no té suficient per extreure patrons. Segurament això vol dir que tot i que és útil, el model el que necessita és més informació sobre la pròpia paraula.

2 Només amb la característica `lemma_pos_tags`

	Idioma	Recall	Precision	F1-Score
BIO	Espanyol	0.5966	0.6400	0.6176
	Neerlandès	0.3895	0.7023	0.5011
BIOES	Espanyol	0.6190	0.6542	0.6361
	Neerlandès	0.4251	0.7254	0.5360

Table 3: Mètriques per al model amb `lemma_pos_tags`

Aquesta característica millora significativament la precisió del model, amb un 10% respecte l'anterior. També s'observa que la codificació BIOES proporciona millors resultats. En aquest cas el neerlandès té una millor precisió. Tot i així, cal tenir en compte el temps d'execució, que augmenta significativament.

3 Totes les característiques menys les dues anteriors (`prefix_suffix, morphology, length, position`)

	Idioma	Recall	Precision	F1-Score
BIO	Espanyol	0.6911	0.7246	0.7074
	Neerlandès	0.6682	0.7182	0.6923
BIOES	Espanyol	0.7109	0.7241	0.7174
	Neerlandès	0.6904	0.7268	0.7081

Table 4: Mètriques per al model amb `prefix_suffix, morphology, length, position`

En aquest cas, el model millora en l'idioma espanyol, igualant-se amb la precisió del neerlandès. No hi ha gaire diferència entre les dues codificacions BIO i BIOES. Pel que fa al f1-score, una mesura més representativa que la presició, veiem que aquest model és molt millors als altres dos anteriors, superant a la feature_function predeterminada. Això ens pot indicar que l'ús combinat de característiques múltiples com prefixos, sufixos, morfologia, longitud i posició en la frase ofereix una aproximació més robusta per a la detecció i el reconeixement d'entitats. Aquesta millora es manifesta tant en el rendiment global del model com en la seva capacitat per adaptar-se i generalitzar millor en ambdós idiomes. Com abans, no hi ha diferències significatives entre les codificacions BIO i BIOES.

4 `context i lemma_pos_tags` (1 i 2)

	Idioma	Recall	Precision	F1-Score
BIO	Espanyol	0.6688	0.7152	0.6912
	Neerlandès	0.5505	0.6590	0.5999
BIOES	Espanyol	0.6774	0.7217	0.6988
	Neerlandès	0.5600	0.6656	0.6083

Table 5: Mètriques per al model amb `context i lemma_pos_tags`

S'ha provat de combinar el primer i el segon experiment. Els resultats mostren que en el cas de l'espanyol, la combinació de les característiques `context` i `lemma_pos_tags` ha portat a una millora en les mètriques de rendiment en comparació amb els experiments anteriors. Però, en el cas del neerlandès, s'observa que el model només amb la característica `lemma_pos_tags` té un rendiment lleugerament millor que quan s'afegeix la característica addicional `context`. En amdós idiomes la codificació BIOS proporciona millors resultats. En conclusió, obtenir el context i el postag ajuda més al model a trobar patrons per l'espanyol que per al neerlandès.

5 `context i prefix_suffix, morphology, length, position` (1 i 3)

	Idioma	Recall	Precision	F1-Score
BIO	Espanyol	0.7270	0.7486	0.7376
	Neerlandès	0.6972	0.7466	0.7211
BIOES	Espanyol	0.7305	0.7537	0.7419
	Neerlandès	0.7187	0.7568	0.7373

Table 6: Mètriques per al model amb `context i prefix_suffix, morphology, length, position`

Amb aquesta configuració de característiques, la codificació BIOES té un rendiment millor en els dos idiomes en comparació amb la BIO. A més, s'observa una millora respecte l'anterior combinació així com una millora respecte els models amb aquestes característiques per separat.

6 `lemma_pos_tags i prefix_suffix, morphology, length, position` (2 i 3)

	Idioma	Recall	Precision	F1-Score
BIO	Espanyol	0.6969	0.7315	0.7137
	Neerlandès	0.6739	0.7240	0.6981
BIOES	Espanyol	0.7109	0.7278	0.7193
	Neerlandès	0.7018	0.7371	0.7190

Table 7: Mètriques per al model amb `lemma_pos_tags i prefix_suffix, morphology, length, position`

El resultats obtinguts són lleugerament millors que amb les característiques per separat, però, tot i proporcionar bons valors de precisió, el temps d'execució penalitza l'eficiència d'aquesta configuració de característiques.

7 Totes les característiques (1, 2 i 3)

	Idioma	Recall	Precision	F1-Score
BIO	Espanyol	0.7357	0.7558	0.7456
	Neerlandès	0.7106	0.7594	0.7342
BIOES	Espanyol	0.7403	0.7622	0.7511
	Neerlandès	0.7190	0.7569	0.7375

Table 8: Mètriques per al model amb totes les característiques addicionals

Finalment, amb la combinació dels tres primers experiments s'obtenen els millors resultats. En aquest experiment, l'espanyol obté millor resultat amb la codificació BIOES, com fins ara, mentre que el neerlandès millora amb la BIO.

2.3 Model final

Un cop s'han escollit els valors de tots els hipeparàmetres i les funcions que s'utilitzaran amb la partició de validació, s'executa el model per a la partició de prova per a veure el seu rendiment.

Finalment, s'ha optat per la configuració de l'experiment 5, ja que s'ha pogut concloure que la característica `lemma_pos_tags` no aporta pràcticament informació i afegeix un augment significatiu i innecessari de temps de còmput durant l'execució del model.

Per tant, tant per a l'idioma espanyol com el neerlandès s'utilitza la codificació BIOES amb totes les carecterístiques menys l'esmentada. A continuació s'observen els resultats obtinguts:

Idioma	Recall	Precision	F1-Score
Espanyol	0.7763	0.7845	0.7804
Neerlandès	0.7628	0.8057	0.7836

Table 9: Mètriques dels models finals

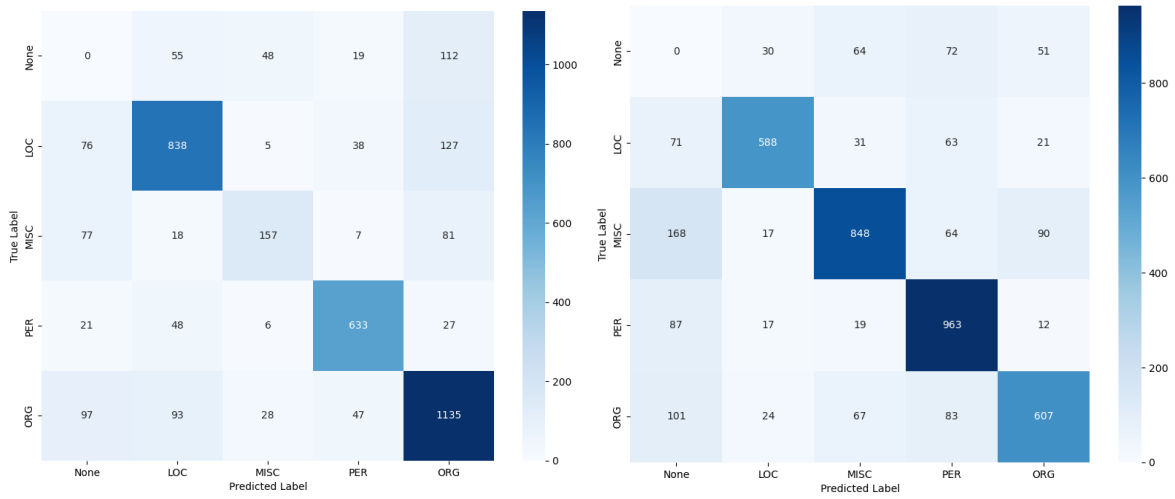

Figure 1: Matriu de confusió dels models finals de l'espanyol i del neerlandès

Els resultats per a l'idioma espanyol mostren un f1-score del 78.04%. Si s'observa la matriu de confusió, es veu que la majoria d'entitats són classificades correctament. La que té un nombre més gran d'errors és l'entitat 'ORG', però això es deu al fet que té un nombre més gran d'entitats en el text, i per tant el model és més propens a classificar una entitat com a organització.

Per a l'idioma neerlandès s'observa una precisió un pèl més alta, arribant al 80% i un f1-score també més alt, amb un 78.36%. En aquest cas, l'entitat amb més errors és 'MISC', a causa del mateix motiu que en l'espanyol. En aquest cas, el model a classificat moltes mostres que no eren cap entitat, com a MISC.

2.4 Model amb textos reals

A part d'haver avaluat els nostres models amb una partició de validació i una de test, també s'ha decidit probar utilitzar el model amb frases reals aleatòries.

Per realitzar aquesta feina, se li ha demanat a una inteligència artificial que ens faci frases que tinguin diferents noms de persones, de llocs i organitzacions. Això s'ha fet ja que no sabem neerlandès i per tant no sabíem com escriure frases per probar el nostre model. S'ha probat amb 10 frases diferents de cada idioma.

En el cas de l'espanyol, en els textos tenim persones i llocs com Lionel Messi, o Paris. El model ha estat capaç de predir les etiquetes amb un F1 de 0.73, així que el model també ha funcionat per aquestes oracions. A la matriu de confusió podem veure que la que pitjor s'ha classificat ha estat les organitzacions, en les que ha fallat en més de la

meitat dels casos. En canvi, per a les localitzacions i persones, té un bon resultat d'encert.

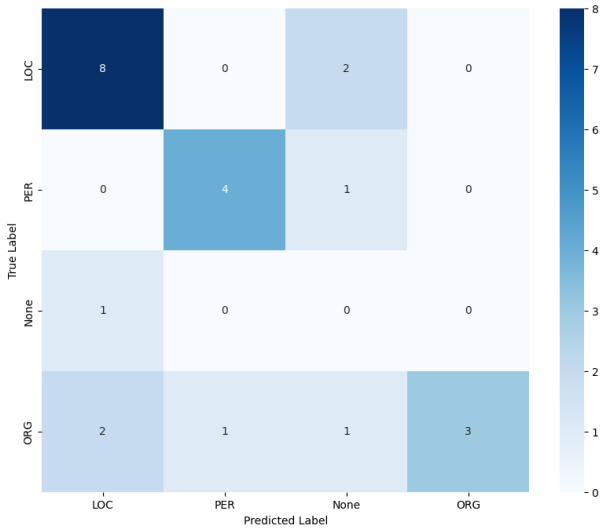

Figure 2: Matriu de confusió de les frases reals en castellà

En el cas del neerlandès, el model també funciona amb un f1 score quasi igual que l'anterior, amb un 0.70. A la matriu de confusió es veu que la majoria d'entitats eren localitzacions. Gairebé totes les localitzacions del text han estat classificades com a tals. En canvi, en el cas de les organitzacions no n'ha classificat cap correctament.

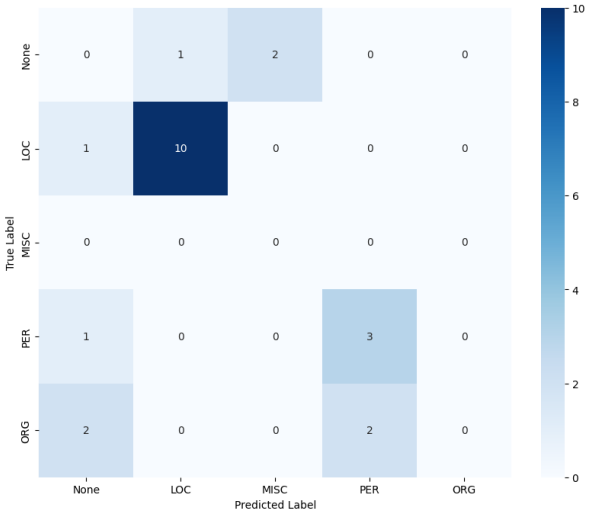

Figure 3: Matriu de confusió de les frases reals en neerlandès

Es podria concloure que el que pitjor prediu el nostre model són les organitzacions. Per solucionar això, es podria fer una ampliació de les funcions de característiques amb Gazetteers o llistes de paraules.

3 Part opcional

Per aquesta segona secció s'ha de fer la mateixa tasca que anteriorment però amb un altre conjunt de dades. Utilitzem el Corpus CADEC (Corpus for Adverse Drug Event Classificatin) que és una base de dades anotada que es centra en els esdeveniments adversos relacionats amb medicaments. Es fa servir en la investigació biomèdica i en el camp de la salut en general pel reconeixement d'entitats anomenades (NER). Tenim les dades dividides en train i test.

Els textos contenen les següents categories d'entitats:

- ADR (Adverse Drug Reaction): Reaccions adverses específiques provocades per un medicament.
- Disease (Di): Malalties o condicions mèdiques mencionades en el text.
- Drug (Dr): Medicaments o substàncies actives mencionats.
- Symptom (S): Símptomes relatats que poden o no estar directament relacionats amb un ADR.
- Finding (F): Troballes clíniques o observacions que no es classifiquen clarament com a símptomes o ADR. Aplicacions de NER amb CADEC

Aquestes dades venien en un format una mica diferent que les del primer corpus, per tant, per utilitzar la nostra classe anterior i entrenar el model s'ha hagut de fer una mica de preprocessament. Les dades estan en forma de matriu, on cada fila es una paraula i cada columna un tipus d'entitat. La nostra funció anomenada `carregar_dataset_opcional` llegeix aquestes dades i les pasa al format demanat: una llista de llistes amb tuples on el primer element és la paraula i el segon és, en BIO encoding, el tipus de entitat (ex. 'B-Dr').

Un cop les dades estan en el format desitjat, ja podem entrenar el model de l'apartat anterior.

3.1 Resultats

A continuació es fa l'anàlisi dels resultats de l'entrenament amb aquestes dades amb els diferents encodings i amb diferents feature functions.

Esquema	Recall	Precisió	F1-Score
BIO	0.4445	0.6283	0.5207
BIOES	0.4257	0.6597	0.5450
IO	0.4257	0.6313	0.5085

Table 10: Mètriques per al model sense feature_functions

A la taula es poden comparar els models sense afegir cap feature_function nova pels tres encodings diferents. Veiem que per aquest dataset és més complicat fer la identificació de entitats solament utilitzant les feature_functions de la classe. Com hem vist en

el apartat anterior, el que ens dona millors resultats és la codificació BIOES, ja que dona més informació i per tant el model és capaç d'aprendre millor els patrons de les dades, tot i donar un f1-score bastant baix.

Com s'ha fet abans, escollim la codificació BIOES ja que és la que més informació ens aporta per crear els patrons del model. Aleshores, ara farem una comparació del rendiment del model amb diferents adicions a la nostra funció de característiques.

Features afegides	Recall	Precisió	F1-Score
Cap	0.4257	0.6597	0.5450
Totes	0.5617	0.6742	0.6128
Totes menys Lemas/Postag	0.5675	0.6894	0.6225
Només Lemas/Postag	0.4675	0.6587	0.5468
Totes menys context	0.5253	0.6473	0.5800

Table 11: Mètriques per al model sense feature_functions

A la taula veiem els resultats del funcionament del model amb les diferents combinacions de adicions a la funció de característiques. Hi ha algunes combinacions que no s'han incluit a la taula ja que no donen bons resultats, com per exemple utilitzant només el context.

Veiem que el model que utilitza totes i totes menys els lemes i el postag són els que tenen millor F1-score. Veiem que els dos superen el 60%. Escollim el model que té totes menys el lema i postag, de manera molt similar al de l'apartat anterior.

4 Conclusions

El projecte actual ha estat capaç d'implementar i evaluar un model de reconeixement d'entitats anomenades (NER) utilitzant diferents esquemes de codificació i funcions de característiques per a les llengües espanyola i neerlandesa. Hem vist que la nostra classe és capaç d'adaptar-se a variacions en les dades i en les funcions de característiques, i hem pogut veure una àmplia visió de com els elements i les dades influencien al rendiment del model.

Pel que fa a les diferents codificacions, s'ha conclòs que és útil utilitzar una que tingui l'inici de la paraula i que sàpiga quan una entitat està a una sola paraula.

En el cas de les funcions de característiques també hem pogut extreure conclusions. La inclusió del context del token (l'anterior i el següent) s'ha vist que és útil però no suficient per si sola per extreure patrons complexos, indicant que el model necessita dades més detallades sobre la pròpia paraula per millorar el rendiment. Afegir les etiquetes POS i els lemes ens ajuda a captar els contextos lingüístics on la forma de les paraules pot variar molt amb la conjugació i declinació. En el notre cas hem conclòs que és la característica que menys ens ha aportat. Finalment, les característiques morfològiques han estat molt útils ja que molts cops són indicatives de certes categories (una majúscula por indicar un nom propi).

Combinar totes aquestes característiques ha resultat en un model robust. La diversitat d'informació ajuda al model a aprendre i generalitzar més.

Els resultats obtinguts reafirmen la importància de l'elecció de funcions de característiques adequades per a l'optimització de models NER. Per futures investigacions, seria útil explorar la integració de noves fonts de dades, com embeddings de paraules contextualitzats o llistes de paraules i gazetteers, que poden capturar contextos més amplis i nuances subtils del llenguatge que van més enllà de l'anàlisi morfològica o sintàctica superficial.

En conclusió, aquest treball ens ha ensenyat que l'ús de CRFs és una bona eina per al reconeixement d'entitats anomenades en múltiples idiomes.