Facultat d'Informàtica de Barcelona

FIB UPC
Processament del Llenguatge Humà

Pràctica 3: Named Entity Recognition

Alumnes :
Granja Bayot, Jordi
Jerez Cubero, Alberto

Tutors :
Medina, Salvador
Turmó, Jordi

April 30, 2024

Contents

1 Introducció

El present informe constitueix la justificació de la solució proposada pel problema plantejat. Aquest és, programar un classificador d'entitats anomenades. Concretament, localitat (LOC), persona (PER), organització (ORG) i altres (MISC), seran les classes a predir, determinades pel còrpora d'ús `CoNLL2002`, tant en Espanyol com en Neerlandès. En aquest sentit, per cada idioma farem servir un model estocàstic discriminatiu, un *Conditional Random Field*, el qual s'utilitza freqüentment per etiquetar dades. D'aquesta manera, podrem establir una comparació entre els dos models, relacionant el preprocessat particular de les dades de cada un amb el seu rendiment.

Entre els objectius, destaca la introducció de conceptes i eines de *Natural Language Processing*, així com l'anàlisi de quins tractaments de les dades s'adeqüen més a la tasca de reconèixer entitats. A més, la comprensió de conceptes relacionats amb els CRF, a través d'un estudi minuciós de les seves opcions de entrenament.

2 Preprocessat de les Dades

En el context particular d'entrenament d'un classificador amb ML, el preprocessament de les dades que duguem a terme té un impacte directe sobre el rendiment del model. Per tant, ens centrarem en estudiar totes les possibilitats d'exploració, des de les diferents codificacions de les entitats, fins a una cerca exhaustiva del millor extractor de *features* de les dades d'entrenament.

2.1 Codificació de les Anotacions

En aquest apartat, ens qüestionem la importància de la codificació dels *tokens* sobre el rendiment dels models. En el context d'etiquetatge d'entitats, el format *Beginning-Inside-Outside* (BIO) és el més popular, si bé n'hi ha d'altres més o menys informatius. En el nostre cas, explorarem totes les codificacions possibles, analitzant quina funciona millor en el context de reconeixement d'entitats.

En aquest sentit, ens cal definir una funció que s'encarregui de la conversió de les dades importades als diferents formats. Aquesta és, la funció `convert_BIO` que, a banda de dur a terme les recodificacions, permet reduir les tuples originals (tok, POS_tag, label) a (tok, label), que és el que el classificador `CRFTagger` realment espera. Pel que fa als formats, podrem treballar amb una simplificació de BIO, que indicarà si una paraula forma part d'una entitat o no (IO), o bé amb ampliacions com BIOS o BIOES, que incorporen la classe *Single*, per indicar entitats d'un únic *token*, i l'etiqueta *End*, quan es tracta d'una paraula que és final d'entitat, simultàniament.

2.2 Extracció de Característiques

En aquest apartat, discutirem el treball pel que respecta a l'extracció de característiques de les dades. Aquesta extracció ha tingut lloc tant en l'àmbit de frase com en l'àmbit de token, on també hem introduït context.

Per a dur a terme aquest procés, cal passar una funció que extregui les característiques d'un *token*, segons dicta el codi font del *CRFTagger* de *NLTK*. En el nostre cas, però, hem optat per la implementació d'una classe amb un mètode *__call__*, que proposa una millora de la funció per defecte *_get_fetures* del *CRFTagger*. La motivació d'aquesta implementació es fonamenta en la capacitat de fer *caching*, a partir dels resultats de l'aplicació del model *spaCy* a cada frase. D'aquesta forma, evitem aplicar el model d'extracció de característiques a cada token d'una frase. En aquest sentit, tenim la capacitat de capturar característiques sensibles al context derivades de l'anàlisi completa de la frase, però amb la necessitat de passar aquesta frase només un cop pel model de spaCy. A més a més, la decisió d'implementar una classe envers una funció, ens brinda la capacitat d'escollir quines característiques extreure en aquella instància del model a través del mètode *__init__*. Això ens serà de gran ajuda de cara al procés de selecció de *features*, com també ho seria de cara a l'ús d'un possible usuari.

Sintetizant l'anterior, al crear una instància de `Feature_getter`, es processa el següent: inicialització de booleans indicant les *features* a extreure, com també de la variable que realitzarà el *caching*; i càrrega del model de *spaCy* corresponent al paràmetre *language*. Tot idioma fora d'espanyol i neerlandès resultarà en una excepció.

Un cop establerta la forma general de la classe, és important descriure les característiques que extraiem de cada token:

- **Context:** Hem introduït totes les característiques del **token anterior** i el **posterior**, codificades amb un prefix de -1_ o +1_, respectivament.
- **Part of Speech (POS):** El POS del token, segons el context que l'envolta, pot ser útil per determinar entitats. A més, sembla especialment rellevant quan interactua amb els tokens anterior i posterior.
- **Lema:** El lema de la paraula ens permet extreure informació sobre el concepte subjacent, cosa que podria ajudar en la predicció.
- **Prefix:** Continuant amb la línia d'extracció de característiques de la classe de *NLTK*, i considerant la seva potencial utilitat, introduïm els tres primers caràcters de la paraula en qüestió.
- **Morfologia:** L'addició de característiques morfològiques obre la porta a una gran varietat de possibilitats. Amb l'objectiu de mantenir baixa la dimensionalitat i afegir informació morfològica rellevant, ens hem centrat en dues característiques: **gènere** i **nombre**.
- **Forma:** A partir d'una paraula, s'extreu la distribució de lletres capitalitzades. Per exemple, del *token* "Hey" s'extrauria "Xxx".

- **Capitalització**.
- **Presència de números**.
- **Símbols de puntuació**.
- **Sufix**.
- **Paraula**.
- **Longitud**.

Aquesta exhaustiva llista de característiques ens proporciona una base completa per al nostre model, permetent-nos capturar una ampla gamma d'informació rellevant per a l'anàlisi i la predicció.

3 Avaluació del rendiment

En aquest apartat, abordem la problemàtica que s'esdevé d'haver de comparar models amb diferents codificacions i, per tant, diferent nombre de classes predites. A més, l'objectiu dels models no és performar bé en termes de les etiquetes BIO, sinó a partir d'aquestes reconèixer correctament les entitats. Per tant, haurem d'establir una mètrica d'avaluació comuna a tots els formats d'etiquetatge, que podem inferir tindrà relació amb les entitats identificades per cada model.

3.1 Mètriques d'avaluació

Com s'ha mencionat prèviament, per a l'avaluació dels models requerim una mètrica comuna que desatengui les particularitats de les diferents codificacions. En el nostre cas, hem considerat oportú adoptar el sistema d'avaluació dissenyat per la cinquena edició de la *Message Understanding Conference* (MUC-5) [1].

Pel sistema en qüestió, s'ignora l'etiqueta *Outside* i, a partir del nivell d'alineació de les etiquetes reals amb les predites, es classifiquen les prediccions en cinc categories.

- **Correct:** l'entitat es correspon exactament amb la real.
- **Partial:** el tipus d'entitat és correcte, però els seus límits no exactament.
- **Incorrect:** els límits són equivalents o quasi, però el tipus d'entitat és erroni.
- **Spurious:** l'entitat no forma part de les dades reals.
- **Misssing:** s'omet completament una entitat real.

Després, amb els comptatges de cada categoria, es defineixen les mètriques que ens serviran per fer comparatives: *precision* i *recall*.

$$Precision = \frac{TP}{TP + FP} = \frac{correct + C \cdot partial}{correct + partial + incorrect + spurious} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} = \frac{correct + C \cdot partial}{correct + partial + incorrect + missing} \tag{2}$$

En ambdues equacions introduïm el terme 'C', que simbolitza la importància que es vol donar a les entitats parcialment encertades. Evidentment, aquest paràmetre haurà de valer menys que 1, per no agafar més pes que les entitats encertades exactament. En el nostre cas, considerem oportú que cada classificació parcial compti 0.5 a la puntuació dels TP del model.

En aquest punt, amb el fi de guanyar perspectiva envers el rendiment real del model, podem relacionar *precision* i *recall* amb la mètrica *F1-score*, ja que totes dues per separat poden conduir a errors d'interpretació. Aquesta serà, la nostra mètrica d'avaluació de referència.

$$F1\_score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3}$$

3.1.1 Implementació de la Funció d'Avaluació

En aquesta secció, es resumeixen els detalls de la implementació de les mètriques exposades a la Sec. 3.1, que s'extrauen a través de la funció `compute_metrics`, dissenyada *ad hoc* per treballar directament amb la sortida dels models. Per aquest motiu i, tenint en compte que la lògica de les mètriques s'articula sobre les entitats, la funció treballa amb un extractor d'entitats `collect_ne`.

Considerant que les nostres mètriques d'avaluació suposen que les entitats no són paraules aïllades, sinó cadenes de *tokens* (e.g. La Coruña), requerim idear una estratègia que tingui presents els límits de les entitats. Per aquest motiu, trobem oportú que el nostre extractor d'entitats treballi amb tuples (ent_type, start_offset, end_offset). D'aquesta manera, un cop recol·lectades les entitats de les dades reals i predites, podrem comparar tant les entitats trobades com els seus límits. Noti's que la sortida de l'extractor assigna una llista d'entitats a cada frase de l'entrada, cosa que computacionalment facilitarà l'avaluació, ja que només haurem de comparar les entitats predites i reals de les mateixes frases. A més, l'extractor d'entitats funciona per qualsevol format de les etiquetes, ja que les discerneix a partir de la classe 'O', present a totes les codificacions.

En aquest punt, desgranades totes les parts bàsiques de la funció d'avaluació, només queda combinar-les. En el nostre cas, donades les frases anotades per *tokens*, tant reals com predites, s'extreuen les entitats, les quals es classifiquen segons les categories exposades a la Sec. 3.1. Finalment, es combinen els comptatges en funció de les Eq's. 1 i 2, i es retornen els resultats totals, així com agregats per tipus d'entitat.

Avaluació amb etiquetes

Alternativament, es defineix una funció secundària per visualitzar els resultats en termes d'etiquetes, és a dir, de la codificació dels *tokens* (BIO). Aquesta és, `bio_classification_report`, la qual farem servir per analitzar més a fons els punts dèbils dels models.

4 Entrenament del Model CRF

4.1 Feature Selection

En aquest apartat, analitzarem l'aportació de l'extracció de les diferents característiques per cada idioma, implementades per la classe `Feature_getter`. La decisió de quedar-nos amb les variables, es fonamentarà en les mètriques d'avaluació esmentades anteriorment, estudiant els seus efectes de manera isolada i amb interacció amb el context. Concretament, prioritzarem les mètriques que treballen sobre entitats, ja que és l'objectiu del treball, més que no en les etiquetes BIO.

4.1.1 Espanyol

Per realitzar la selecció de característiques, hem seguit un procés iteratiu. En una primera instància, vam emmagatzemar els resultats d'un model base, resultant de la funció d'extracció per defecte del *CRFTagger*. Després, vam provar d'introduir també característiques relatives al context. Finalment, vam explorar l'aportació de les nostres variables addicionals, incloent-hi i excloent informació contextual dels *tokens*.

La taula de resultats sobre el conjunt de validació és la següent:

	total acc	total recall	**total F1**
Baseline	0.707	0.666	**0.686**
Prev_tok	0.746	0.716	**0.731**
Prev_tok_Next	0.756	0.733	**0.745**
Baseline_wMorpho	0.706	0.668	**0.686**
Prev_tok_wMorpho	0.747	0.722	**0.734**
Prev_tok_Next_wMorpho	0.750	0.728	**0.739**
Baseline_wAll	0.719	0.691	**0.705**
Prev_tok_wAll	0.748	0.733	**0.740**
Prev_tok_Next_wAll	0.753	0.739	**0.746**

Table 1: Resultats de l'avaluació del sistema en mètriques generals.

	PER F1	ORG F1	LOC F1	MISC F1
Baseline	0.769	0.728	0.588	0.573
Prev_tok	0.809	0.774	0.653	0.558
Prev_tok_Next	0.837	0.804	0.642	0.544
Baseline_wMorpho	0.772	0.726	0.588	0.570
Prev_tok_wMorpho	0.821	0.781	0.647	0.555
Prev_tok_Next_wMorpho	0.841	0.797	0.631	0.535
Baseline_wAll	0.805	0.756	0.591	0.577
Prev_tok_wAll	0.850	0.790	0.641	0.551
Prev_tok_Next_wAll	0.858	0.799	0.636	0.560

Table 2: Resultats de l'avaluació agrupada per entitats.

En la taula 1, es pot observar una millora dràstica en la mètrica F1 amb la introducció de context. D'altra banda, es nota que la morfologia no té un impacte significatiu en el rendiment, mentre que l'addició de les característiques que faltaven millora lleugerament els resultats. Pel que fa a la taula 2, podem veure com la millora es nota majoritàriament en la identificació de persones. També podem veure canvis notables en les classes d'organització i localització. Per contra, els models tenen dificultat per predir la classe MISC. Més endavant, analitzarem quines variables ajuden a discernir cada classe, a l'apartat de *feature importance*

En vista d'això, la nostra solució proposa quedar-se amb l'últim model, el qual implementa totes les característiques proposades.

4.1.2 Neerlandès

Anàl·logament, fem la selecció de variables pel model en neerlandès, obtenint els següents resultats:

	total acc	total recall	**total F1**
Baseline	0.649	0.586	**0.616**
Prev_tok	0.721	0.663	**0.690**
Prev_tok_Next	0.730	0.669	**0.698**
Baseline_wMorpho	0.648	0.604	**0.625**
Prev_tok_wMorpho	0.721	0.678	**0.699**
Prev_tok_Next_wMorpho	0.721	0.673	**0.696**
Baseline_wAll	0.718	0.699	**0.708**
Prev_tok_wAll	0.772	0.743	**0.757**
Prev_tok_Next_wAll	0.778	0.748	**0.762**

Table 3: Resultats de l'avaluació del sistema en mètriques generals.

	PER F1	ORG F1	LOC F1	MISC F1
Baseline	0.572	0.736	0.652	0.574
Prev_tok	0.639	0.759	0.728	0.681
Prev_tok_Next	0.645	0.774	0.729	0.689
Baseline_wMorpho	0.574	0.770	0.631	0.599
Prev_tok_wMorpho	0.660	0.790	0.688	0.695
Prev_tok_Next_wMorpho	0.646	0.769	0.712	0.695
Baseline_wAll	0.644	0.806	0.732	0.703
Prev_tok_wAll	0.705	0.798	0.772	0.783
Prev_tok_Next_wAll	0.726	0.799	0.767	0.790

Table 4: Resultats de l'avaluació agrupada per entitats.

En aquest cas, l'anàlisi dels resultats és molt similar al del model en espanyol. No obstant, les puntuacions de F1 en termes d'entitats difereixen. A diferència del model en espanyol, observem com el model complet prediu notablement totes les classes. Això ens indica

que, pel neerlandès, les nostres *features* permeten no marginar cap entitat, resultant en millors models.

Quant a la selecció de variables, s'ha optat per incorporar totes les *features* de cada *token*, així com les del seu antecessor. A diferència del model en espanyol, no s'han considerat les característiques del *token* predecessor. D'aquesta manera, podrem estudiar com afecta entrenar dos models amb dimensionalitats diferents, tant a nivell de rendiment com de computació.

4.2 Optimització d'hiperparàmetres

Un cop realitzada la selecció de variables, detallarem el procediment que s'ha dut a terme per a optimitzar els hiperparàmetres dels models. Segons la documentació, els que semblen afectar més el rendiment de les prediccions de la classe *CRFTagger* són 'c1', 'c2', 'max_iterations', 'minfreq', 'possible_states' i 'possible_transitions'. En aquest sentit, per tal de reduir la combinatòria de models possibles, s'ha considerat oportú dividir la cerca en dues passes. Primer, trobar els millors paràmetres per 'c1', 'c2' i 'max_iterations', en considerar-se els més importants, i després buscar la resta. Tot i no resultar en un resultat òptim, serà una bona aproximació.

Novament, la cerca es fa de forma separada, és a dir, obtindrem uns hiperparàmetres diferents per cada idioma. Això es fa així per preservar les particularitats de cada llenguatge. No obstant, l'espai de búsqueda és el mateix i es basa en l'estat de l'art (e.g. potències de deu pels regularitzadors). Seguint aquest procediment, s'obtenen els següents resultats:

	c1	c2	max_iter	poss_transitions	poss_states	min_freq	F1
Espanyol	0.01	1	200	False	True	0	**0.756**
Neerlandès	0.01	0.1	50	True	True	0	**0.776**

Table 5: Hiperparàmetres òptims per a l'entrenament de cada model.

Com podem observar en la taula 5, l'optimització dels hiperparàmetres resulta en un increment de F1 per la partició de validació. Per tant, mantindrem aquesta informació pel procés d'avaluació dels models, amb l'esperança d'obtenir els millors resultats possibles.

Cal destacar que, a causa de la dimensionalitat de cada model, la cerca en el cas de l'espanyol ha costat bastant més computacionalment. Llavors, aquest apartat, en què s'han hagut d'entrenar múltiples models, ens fa qüestionar si realment val la pena assumir més dimensionalitat quan la millora en rendiment és nímia. En aquest cas, intentem trobar els millors models i els recursos computacionals són suficients però, en problemes més escalables, optaríem per models més simples.

5 Anàlisi dels Resultats

En aquesta secció, valorem els resultats dels nostres esforços. Després de detallar les mètriques d'avaluació i trobar les variables i hiperparàmetres que les optimitzen, veurem com els models es comporten davant dades no vistes.

5.1 Neerlandès

Class	Precision	Recall	F1-score	Support
B-LOC	0.87	0.81	0.84	774
I-LOC	0.67	0.53	0.59	49
B-MISC	0.87	0.76	0.81	1187
I-MISC	0.63	0.46	0.53	410
B-ORG	0.81	0.71	0.76	882
I-ORG	0.80	0.65	0.72	551
B-PER	0.78	0.90	0.83	1098
I-PER	0.87	0.96	0.91	807

Table 6: Avaluació a nivell de codificació del model en neerlandès.

Metric	Acc	Recall	Total F1	PER F1	ORG F1	LOC F1	MISC F1
Dutch_BIO	0.814	0.789	**0.801**	0.777	0.769	0.867	0.811

Table 7: Avaluació a nivell d'entitats del model en neerlandès.

Quant al model BIO en neerlandès, trobem una puntuació F1 satisfactòria. Si analitzem els punts dèbils, podem destacar un rendient molt baix en les classes *Inside* de les entitats *MISC* i *LOC*, segurament per la manca de mostres d'aquests tipus a les dades d'entrenament. No obstant, són aquestes entitats les que millor F1 reporten (taula 7), a diferència d'altres amb millors rendiments a les seves respectives classes *Inside*. Això, justifica la nostra línia de basar les nostres decisions en mètriques a nivell d'entitat. En aquest sentit, ens preguntem si diferents codificacions s'ajusten millor a la tasca de reconèixer entitats en neerlandès.

Model	Acc	Recall	Total F1	PER F1	ORG F1	LOC F1	MISC F1
Dutch_BIO	0.814	0.789	**0.801**	0.777	0.769	0.867	0.811
Dutch_IO	0.792	0.758	**0.775**	0.774	0.741	0.830	0.761
Dutch_BIOS	0.810	0.784	**0.797**	0.775	0.752	0.876	0.804
Dutch_BIOES	0.811	0.787	**0.799**	0.784	0.747	0.874	0.804

Table 8: Avaluació a nivell d'entitats de models amb diferents codificacions.

Com es pot observar en la taula 8, el model més útil per la tasca en qüestió és el derivat de la codificació d'entitats BIO. Segurament, pel fet que els hiperparàmetres s'han ajustat a aquest format en concret. No obstant això, podem extreure que la incorporació de l'etiqueta *Single* resulta en bons models. Això ens indica que les nostres dades presenten entitats d'una sola paraula. D'altra banda, en qualsevol cas la codificació IO és la millor, per la seva manca d'aportació informativa.

En aquest sentit, acabarem l'avaluació amb el model neerlandès BIO. Quantificarem el seu rendiment final i intentarem explicar-lo a partir de les seves variables més importants.

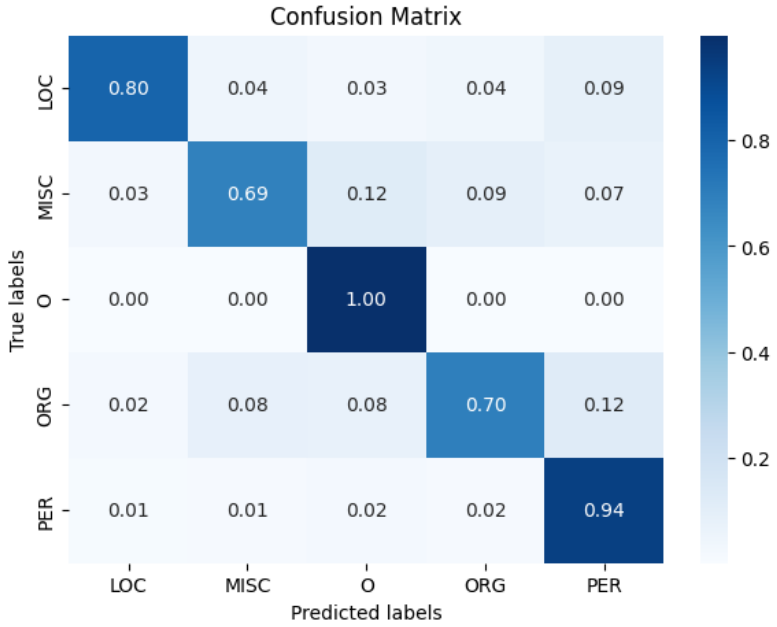

Figure 1: Matriu de confusió del model final en neerlandès.

Com mostra la figura 1, el nostre model encerta totes les etiquetes *Outside*, i gairebé totes les de l'entitat persona. No obstant, en aquesta última classe la puntuació F1 de 0.777 ens indica que el seu comptatge *spurious* ha de ser alt. La matriu de confusió ho confirma: es prediuen com a *PER* gairebé un 10% de les altres entitats reals. Per tant, podem concloure que el nostre model està esbiaixat cap a la classe *PER*. D'altra banda, els seus punts dèbil són les classes *MISC* i *ORG*.

En aquest sentit, si obtenim les variables més importants pel model, trobem gran quantitat associades a les classes *Outside* i *LOC*, destacant per aquesta última les *features* 'WORD_Gent', 'SUF_urs' i 'PRE_VS'. Obivament, també hi ha d'associades a l'entitat persona, com 'PRE_Chr' o 'WORD_Vanderpoorten'. No obstant, no trobem quasi variables associades a la classe *ORG*, fet que explica que sigui un punt dèbil del model. Per contra, no podem dir el mateix de l'entitat *MISC*, en què si hi han variables importants, majoritàriament derivades de la forma (e.g. 'SHAPE_XXX-xxxx'). Per tant, podem concloure que es tracta d'una classe estranya, que presenta un patró difícilment capturable pel model.

5.2 Espanyol

Anàlogament, podem desplegar el mateix estudi pel model en espanyol, amb els seus corresponents hiperparàmetres optimitzats.

Class	Precision	Recall	F1-score	Support
B-LOC	0.80	0.79	0.79	1084
I-LOC	0.65	0.65	0.65	325
B-MISC	0.69	0.52	0.59	339
I-MISC	0.66	0.53	0.59	557
B-ORG	0.81	0.84	0.83	1400
I-ORG	0.83	0.79	0.81	1104
B-PER	0.85	0.89	0.87	735
I-PER	0.90	0.95	0.92	634

Table 9: Avaluació a nivell de codificació del model en espanyol.

Metric	Acc	Recall	Total F1	PER F1	ORG F1	LOC F1	MISC F1
Spanish_BIO	0.797	0.792	**0.795**	0.853	0.796	0.790	0.650

Table 10: Avaluació a nivell d'entitats del model en espanyol.

Respecte al model BIO en espanyol, trobem uns resultats F1 notables però, a diferència del neerlandès, més desequilibrats quant a entitats. Si bé, les puntuacions per a les entitats *PER* i *ORG* són significativament millors. No obstant això, novament la classe *MISC* resulta la més difícil de predir, segurament per la baixa quantitat de mostres d'aquest tipus. En aquest punt, ens preguntem si la codificació BIO és la més apropiada per predir entitats en neerlandès.

Model	Acc	Recall	Total F1	PER F1	ORG F1	LOC F1	MISC F1
Spanish_BIO	0.797	0.792	**0.795**	0.853	0.796	0.790	0.650
Spanish_IO	0.789	0.781	**0.785**	0.822	0.791	0.791	0.632
Spanish_BIOS	0.799	0.794	**0.796**	0.854	0.794	0.801	0.640
Spanish_BIOES	0.799	0.796	**0.797**	0.854	0.791	0.802	0.652

Table 11: Avaluació a nivell d'entitats de models amb diferents codificacions.

Segons la taula 11, el format més adequat de les anotacions és el BIOES, amb una millora considerable en la classe *LOC*. Això ens indica que moltes d'aquestes entitats són *single token*, per la qual cosa es beneficien de la codificació. Per tant, amb el fi d'obtenir el millor model ens quedem amb el format BIOES.

Class	Precision	Recall	F1-score	Support
B-LOC	0.73	0.65	0.69	196
E-LOC	0.75	0.69	0.72	177
I-LOC	0.65	0.58	0.61	148
S-LOC	0.82	0.83	0.82	888
B-MISC	0.66	0.52	0.59	183
E-MISC	0.58	0.48	0.52	183
I-MISC	0.72	0.49	0.58	374
S-MISC	0.70	0.44	0.54	156
B-ORG	0.85	0.78	0.81	467
E-ORG	0.78	0.72	0.75	458
I-ORG	0.81	0.80	0.81	646
S-ORG	0.78	0.87	0.82	933
B-PER	0.88	0.96	0.92	504
E-PER	0.87	0.96	0.91	498
I-PER	0.87	0.82	0.84	136
S-PER	0.77	0.74	0.75	231

Table 12: Avaluació a nivell de codificació del model BIOES en espanyol.

En aquest sentit, podem observar en la taula 12 com les puntuacions F1 de cada classe han millorat. Tot i augmentar el nombre de classes, reduint així el nombre de mostres de cada una, el model és capaç d'extreure les característiques necessàries per distingir-les. No obstant això, aquesta anàlisi confirma la impossibilitat de performar bé en l'entitat *MISC*, com passava també en el cas neerlandès. Per acabar, podem resumir aquesta informació en una matriu de confusió, normalitzada en vers la classe positiva.

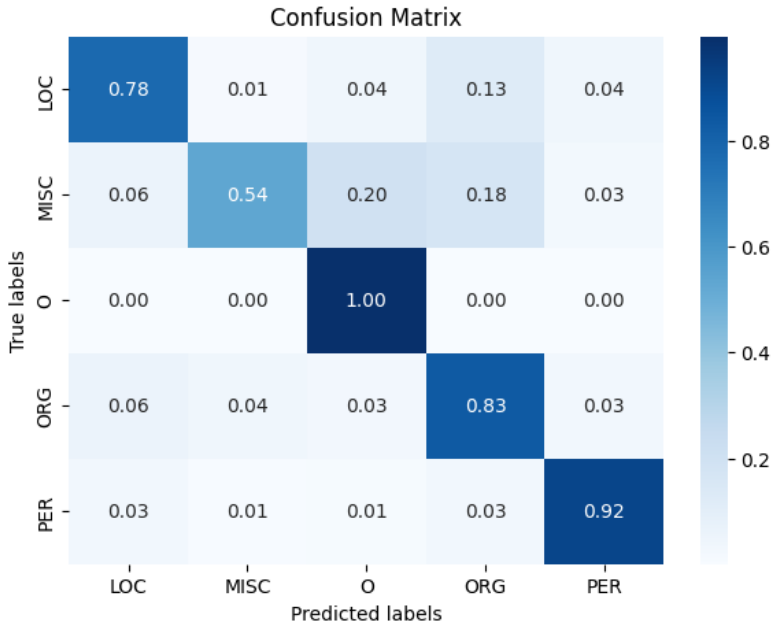

Figure 2: Matriu de confusió del model final en espanyol.

Segons la figura 2 el model final per l'espanyol prediu correctament tots els *Outside tokens*. A més, l'entitat persona és el seu punt fort i, a diferència del cas neerlandès, no podem determinar un biaix cap a aquesta classe. Si bé, podem fer-ho en el cas de *Outside* o *ORG*, que es prediuen sovint en detriment de les classes *MISC* i *LOC*.

A partir d'aquesta anàlisi, podem inferir que les entitats persona en espanyol són més diferenciables de les altres entitats que en el cas del neerlandès. Això coincideix amb la presència de característiques importants per aquesta classe, com 'SHAPE_X' o 'LEN_1', que ens indiquen la presència d'abreviacions (e.g. J. Antoni). D'altra banda, la bona actuació en les entitats *ORG* s'explica amb la presència de variables rellevants com '-1_WORD_.' o '+1_HAS_NUM'. Per contra, no trobem *features* amb alt pes per la classe *LOC*, fet que pot explicar que sigui un punt dèbil del model. Finalment, anàlogament al neerlandès, tot i presenciar variables associades a l'entitat *MISC*, la mala actuació del model ens condueix a pensar que és una classe molt sorollosa i poc freqüent.

6 Opcional: CADEC

En aquest apartat discutirem la implementació i predicció d'entitats en el conjunt opcional **CADEC** (Corpus of Adverse Drug Event Annotations). Aquest conjunt conté informació sobre esdeveniments adversos de drogues (ADE). La base de dades inclou fragments de text anotats que descriuen reaccions adverses causades per fàrmacs.

L'objectiu, un cop més, és l'ús d'un model estocàstic discriminatiu per a classificar i etiquetar entitats anomenades. Un cop més, usarem el model *Conditional Random Fields*.

6.1 Estructura i preprocessament de les dades

En primer lloc, necessitarem un preprocessament per a extreure'n una entrada vàlida per al *CRFTagger*. Així doncs, primer cal mirar l'estructura d'aquest conjunt de dades. Cada fila d'aquest conjunt es basa en sis elements de la forma: **Token**, Adverse Drug Reaction (**ADR**), Disease (**Di**), Drug (**Dr**), Symptom (**S**), Finding (**F**). Ja que a nosaltres ens interessa una entrada (token, label), hem definit una funció que realitza el següent per cada frase:

1. Si la frase és buida, afegirem la frase anterior al conjunt de frases preprocessades.
2. Mirem que no sigui un token que representi el títol del medicament.
3. Per cada element de la frase mirem si alguna d'elles té etiqueta i afegim l'element junt amb l'etiqueta a la frase. Si cap en té, li assignem l'etiqueta '0'.

Un cop passem per aquest procés i per la funció *convert_BIO*, obtindrem les frases amb la forma (token, label) que ens interessa.

6.2 Resultats

Els resultats d'aquest model han estat avaluats per a cada una de les codificacions. Els resultats han estat els següents:

	acc	recall	**total F1**	ADR F1	Di F1	Dr F1	S F1	F F1
BIO_w/o_hiper	0.736	0.606	**0.665**	0.665	0.306	0.847	0.247	0.182
BIO_hiper	0.738	0.609	**0.668**	0.671	0.311	0.847	0.256	0.164
IO_hiper	0.731	0.605	**0.662**	0.659	0.288	0.846	0.254	0.175
BIOS_hiper	0.733	0.613	**0.667**	0.664	0.284	0.852	0.222	0.216
BIOES_hiper	0.730	0.604	**0.661**	0.660	0.235	0.845	0.182	0.216

Table 13: Taula de resultats conunt **CADEC**

La taula 13 ens mostra les mètriques per Named Entities de les diferents codificacions. La accuracy pot semblar bona, ja que sembla poder capturar d'una manera general on hi ha Named Entities. D'altra banda, ens trobem amb una F1 bastant millorable, on destaca la poca variància de resultats respecte al tipus de codificació. Els resultats tan dolents en les columnes com **F**, **S**, **Di** respecte a columnes com **Dr**, junt amb la relativament

alta *accuracy*, ens fan entendre que el classificador no ha après els patrons de les classes minoritàries. Aquesta hipòtesi a priori sembla reforçada quan observem la diferència de la *accuracy* amb el *F1*. Per a poder afirmar això, però, cal observar com es distribueixen les prediccions en una matriu de confusió:

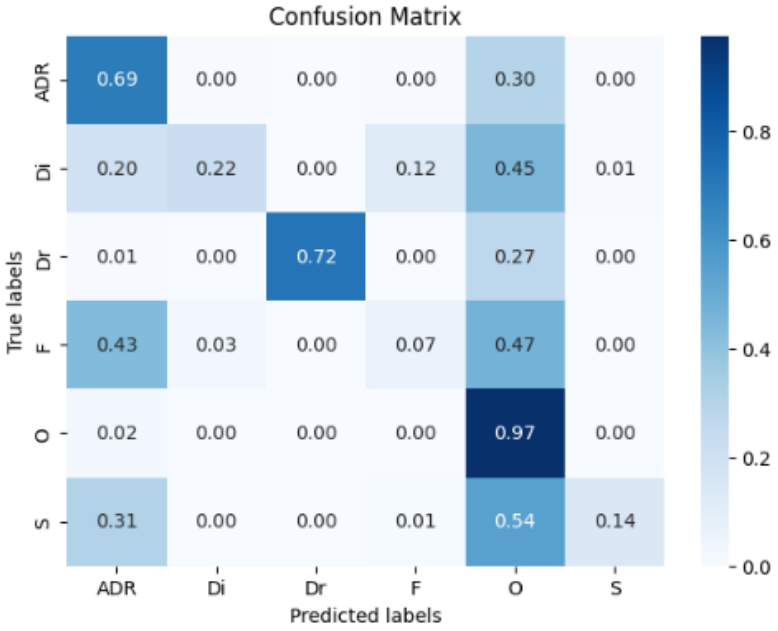

Figure 3: Matriu de confusió conjunt CADEC.

En la figura 3 podem observar el recall a cada cel·la de la matriu de confusió. Més específicament, podem veure que les prediccions que haurien de ser favorables a les tres classes minoritàries **S, Di, F**, han quedat totalment absorbides per l'etiqueta majoritària **O**. Fet que porta al fet que no es prediguin les paraules com a part d'una entitat, i per tant que afecta les mètriques. Això pot trobar sentit en visualitzar el nombre d'etiquetes per classe al conjunt d'entrenament.

Label	Count
O	76199
ADR	12100
Dr	1548
F	600
S	447
Di	355

Table 14: Recompte per classe del conjunt d'entrenament.

Tot i que sembla explicar el comportament, sobta que la diferència de recompte per exemple entre les classes DR i F porti a una diferència tan gran en els resultats.

Després d'analitzar el conjunt de dades, els factors que semblen influir en el rendiment millorable del model són:

1. Discontinuïtat d'algunes entitats al llarg de les respectives frases.
2. Classes amb molt poques instàncies d'entrenament, per tant, amb dificultat per a aprendre els seus patrons, així essent classificades com a no-entitats.
3. Dificultat aparentment inherent en la classificació d'algunes categories com a entitats (com en el cas de la categoria F).

Respecte al primer dels punts, podem afirmar que el *CRFTagger* amb anotacions *BIO* no acaba de ser adequat per la tasca proposada. Això és principalment degut a l'assumpció que les entitats són contínues dins de les frases. Continuïtat no garantida en les anotacions del conjunt de dades.

D'aquesta manera, pensem que de cara a poder millorar el model, seria d'ajuda poder reconèixer els patrons de les classes poc representades perquè passin a ser predites com a entitat, comportament que veiem que no s'està produint de manera adequada. Això ho podríem aconseguir amb més dades o, d'altra banda, aprofitant més la capacitat del model amb un ajustament d'hiperparàmetres més profund, entre altres.

7 Conclusions

En resum, en aquest treball hem pogut introduir conceptes del reconeixement d'entitats nombrades, així com comprendre les seves dificultat associades. D'una banda, hem analitzat els diferents formats de les anotacions dels *tokens*, a partir de la implementació d'una mètrica comuna, basada en la capacitat predictiva dels models a nivell d'entitat.

Pel que fa als models resultants, cal destacar un gran esforç en la selecció de les millors variables i hiperparàmetres, a través de les particions de dades de validació. En aquest sentit, hem pogut observar com tot model es beneficiava de l'addició de característiques en termes de puntuació F1, però el preu a pagar era un cost computacional més alt. Finalment, el model espanyol i neerlandès presentaven resultats similars, però més equilibrats en vers les entitats en aquest últim cas. El que està clar, és que la classe *MISC* en ambdós casos ha davallat el rendiment total, en interpretar-se com una entitat poc representada i difusa. Per aquest motiu, futures accions en el treball, podrien estar encaminades a prescindir d'aquesta classe, enfocant els esforços del model a predir la resta d'entitats, amb significat explícit.

Per acabar, hem provat d'extrapolar la nostra implementació a altres dades, com són les del CADEC. Les seves particularitats, però, han conduït a pitjors resultats, en quant que l'assumpció d'entitats contínues no es segueix. Per tant, s'hauria d'idear una metodologia diferent de l'aplicada pels models anteriors.

References

[1] Nancy Chinchor and Beth Sundheim. 1993. MUC-5 Evaluation Metrics. In Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993.