Informe de la Pràctica 3

Lluc Furriols, Pau Prat

Maig de 2024

Universitat Politècnica de Catalunya
Grau en Intel·ligència Artificial
Processament del Llenguatge Humà

Resum

Aquest document correspon a l'informe de la tercera pràctica de l'assignatura Processament del Llenguatge Humà del Grau en Intel·ligència Artificial de la Universitat Politècnica de Catalunya (UPC).

En aquest informe es descriu el desenvolupament i la validació de models de reconeixement d'entitats i etiquetatge de parts del discurs (POS) utilitzant CRFTagger per a textos en espanyol i neerlandès. El projecte comença amb una descripció detallada de les funcions utilitzades per a processar els textos, així com les funcions per a l'extracció i avaluació d'entitats. Seguidament, es detalla l'entrenament de models per a la predicció de POS amb modificacions en la funció de característiques per obtenir millors resultats. Finalment, s'analitzen els resultats obtinguts en l'execució dels models amb textos reals, indicant com s'han comportat davant textos mai vistos i quines millores podriem realitzar en un futur.

Índex

1 Funcions essencials

En aquest primer apartat, començarem explicant les funcions que utilitzarem al llarg d'aquest projecte.

1.1 get_token, get_token_POS i get_token_entity

Aquestes són les tres primers funcions que utilitzarem. Aquestes funcions reben com a paràmetre una paraula en format: (token, categoria gramatical, entitat), i retornen la part d'aquesta tupla que el seu nom indica. És a dir, `get_token()` retorna una llista només amb els tokens, `get_token_POS()` retorna una llista de tuples (token, POS) i `get_token_entity()`, en canvi, les tuples són (token, entitat).

1.2 extract_entities

La funció rep dos arguments: **tagged_words**, que és una llista de tuples que contenen la paraula del text i la seva etiqueta, i **encoding**, que especifica el tipus de codificació utilitzat per les etiquetes ('BIO'/'IO'/'BIOW').
La funció procedeix a trobar les entitats en les paraules que li han arribat com a paràmetre, tenint en compte la codificaició que se li ha esepecificat. Finalment la funció retorna una llista amb les entitats que hi ha al text d'entrada. És a dir si la funció rep aquesta frase: [(Avui, O), (en, O), (Tomas, B-PER), (Molina, I-PER)], la funció retornarà l'índex inicial i final de totes les entitats, en aquest cas: [((2,3, PER)]. Aquesta funció és útil, per més endavant, quan vulguem avaluar el rendiment dels models, així tindrem una llista amb només les entitats.

1.3 evaluate_entities

Aquesta funció és l'encarregada d'avaluar el rendiment dels nostres models. Rep com a arguments dues llistes: les entitats que ha predit el model, i les entitats que hauria de haver predit (reals). Les llistes que rep la funció, han estat prèviament preprocessades per la funció 1.2. Així doncs, la funció `evaluate_entites` ha de comparar els dos conjunts que ha rebut. Cada tupla conté tres elements que representen l'índex inicial, l'índex final i el tipus de l'entitat dins del text.
Cada entitat és avaluada com un conjunt per permetre una comparació eficient entre les entitats predites i les veritables. Utilitzant conjunts, la funció determina els veritables positius (true positives), els falsos positius (false positives) i els falsos negatius (false negatives):

- **Veritables positius (True Positives)**: Entitats que el model ha predit correctament.
- **Falsos positius (False Positives)**: Entitats que el model ha predit incorrectament com a presents.
- **Falsos negatius (False Negatives)**: Entitats reals que el model no ha detectat.

A partir d'aquests comptes, es calculen les mètriques següents:

- **Precisió (Precision)**: Proporció d'entitats correctament identificades entre totes les entitats identificades per el model.

$$\text{Precisió} = \frac{\text{True Positives}}{\text{True Positives} + \text{True Negatives}} \tag{1}$$

- **Recuperació (Recall)**: Proporció d'entitats correctament identificades entre totes les entitats reals.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

- **Puntuació F1 (F1 Score)**: Mitjana harmònica de la precisió i la recuperació, proporcionant una mesura única de la qualitat del model.

$$\text{F1-score} = 2 \times \frac{\text{Precisió} \times \text{Recall}}{\text{Precisió} + \text{Recall}} \tag{3}$$

Cal destacar l'ús de conjunts per avaluar les entitats reconegudes pel model, el qual ofereix diversos avantatges significatius en el context de la comparació i avaluació de les entitats; permeten operacions de comparació ràpides, especialment quan es tracta de determinar interseccions i diferències, i les entrades duplicades són automàticament eliminades, per tant, d'aquesta manera ens assegurem que cada entitat única sigui comptada només una vegada, evitant augmentar els veritables positius o els falsos positius quan la mateixa entitat apareix diverses vegades.

2 Model per predir POS tag

Hem desenvolupat un model específic per a la predicció de POS tags (Parts of Speech) que classifica cada token en categories gramaticals com a noms, adjectius, verbs, entre altres.

2.1 Desenvolupament del Model

Per millorar la precisió de la predicció de POS tags, vam decidir crear una classe personalitzada, `FeatureGetterPOS`, que substitueix la funció de característiques per defecte del CRFTagger. Aquesta nova classe incorpora diverses característiques avançades:

- **Detecció de Dígits:** Identifica la presència de números dins dels tokens.
- **Detecció de Majúscules:** Detecta si els tokens comencen amb majúscules, una pista útil per identificar noms propis o inicis de frase.
- **Signes de Puntuació:** Reconeix la presència de signes de puntuació, que poden influir en la interpretació gramatical dels tokens adjacents.
- **Anàlisi de Prefixos i Sufixos:** Extreu prefixos i sufixos dels tokens, que són indicatius de la naturalesa gramatical dels tokens.
- **Context:** Inclou informació sobre els tokens anteriors i posteriors, aportant un context addicional que pot ser crucial per determinar la categoria gramatical correcta.

Aquestes característiques són calculades per cada token en el text i s'utilitzen com a entrades per al model CRFTagger que entrenem posteriorment.

2.2 Entrenament i Avaluació del Model

El model ha estat entrenat individualment per a cada idioma (espanyol (esp) i neerlandès (ned)) per assegurar-nos que les particularitats lingüístiques de cada idioma siguin tractades correctament. Els resultats de l'entrenament han estat realment satisfactoris, doncs hem obtingut una precisió del:

- Espanyol (esp): 96.1%
- Neerlandès (ned): 95.8%

2.3 Implementació i Reutilització

Els models entrenats s'han guardat a la mateixa carpeta per facilitar la seva reutilització en futures sessions de processament de text, evitant així la necessitat de reentrenar els models cada vegada, lo qual optimitza molt el temps d'execució, sobretot a l'hora de modificar certes parts del codi, o quan havíem de reiniciar el kernel per una certa raó.

3 Model per predir entitats amb codificació BIO

Per començar, vam entrenar dos models diferents:

El primer va ser amb el CRFTagger sense introduir cap modificació en el procés. L'entrenament es va realitzar utilitzant el següent format: cada parell (token, entity), on 'token' fa referència a la paraula analitzada, mentre que 'entity' representa la classe o categoria d'entitat assignada a aquesta paraula d'acord amb la codificació definida.

Per millorar el rendiment d'aquest primer model, vam optar per introduir diverses modificacions significatives al procés: vam crear una classe que modificava la funció get_features per defecte del CRFTagger. Aquesta nova classe enriquia la informació utilitzada per al reconeixement d'entitats.

La nova classe té la pràcticament la mateixa estructura que la que hem utilitzat abans per predir els POS tags en la secció 2, ja que ha demostrat tenir un bon rendiment. És a dir, les noves funcionalitats incorporades per aquesta classe ampliada abasten la detecció de dígits, majúscules i signes de puntuació. A més, s'analitzen els sufixos i prefixes dels tokens, i s'incorpora una consideració contextual mitjançant la identificació de les paraules anteriors i posteriors. A més a més, aquesta nova funció get_features també obté informació sobre la categoria gramatical, i és obtinguda utilitzant el model entrenat prèviament explicat en la secció 2. És cert que podríem haver aportat inclús més informació però vam considerar que així teníem una bona relació entre calitat i quantitat.

Vam fer aquest procés tant per el idioma espanyol com pel neerlandès. Abans d'entrenar els models respectius, vam crear una funció anomenada grid_search, la qual amb la ajuda d'un protocol de validació, ens va ajudar a triar quina és la millor combinació de característiques. La resolució d'aquesta funció va ser que utilitzant totes les característiques és com millor resultats donava.

Un cop teníem la millor combinació de característiques vam poder entrenar els models.

3.1 Avaluació dels resultats

Per a la valoració del rendiment del model, vam optar per avaluar-lo exclusivament en termes d'entitats reconegudes. Això implica que només considerem les paraules que formen part d'una entitat per a la valoració, excloent-ne les que es troben fora d'aquest context. Aquest enfocament ens permet obtenir mètriques de rendiment que reflecteixen de manera més precisa la capacitat del sistema en la tasca de reconeixement d'entitats.

Amb els models que hem entrenat prèviament, vam dur a terme la predicció de les entitats sobre el conjunt de proves. Simultàniament, vam obtenir les etiquetes reals, que serien utilitzades per avaluar els models. Per a això, vam utilitzar la funció `extract_entities` (veure secció 1.2) per extreure les entitats de les llistes predites i reals. Posteriorment, vam avaluar el rendiment dels models emprant la funció `evaluate_entities` (veure secció 1.3). Aquesta avaluació ens proporciona una visió objectiva de la precisió, el recall i la puntuació F1 dels nostres models en la tasca de reconeixement d'entitats.
Els resultats obtinguts pel model inicial es mostren a continuació:

	CRFTagger Bàsic	**CRFTagger Modificat**
Resultats per Espanyol		
Precisió	0.741	0.783
Recall	0.708	0.768
F1 Score	0.724	0.776
Resultats per Neerlandès		
Precisió	0.701	0.769
Recall	0.621	0.715
F1 Score	0.659	0.741

Taula 1: Comparació del rendiment del CRFTagger

3.1.1 Anàlisi resultats

Els resultats obtinguts per als models CRFTagger en els idiomes Espanyol i Neerlandès mostren una millora en les mètriques de rendiment amb la implementació de les modificacions.

Per l'idioma Espanyol, el model CRFTagger modificat mostra millores significatives en totes les mètriques respecte al model bàsic.

Pel que fa a l'idioma Neerlandès, també s'observa una millora en les mètriques amb la implementació de les modificacions.

En general, els resultats mostren que les modificacions aplicades als models CRFTagger han estat efectives en la millora del seu rendiment en la tasca de reconeixement d'entitats, tant per a l'idioma Espanyol com per al Neerlandès. Això suggereix que les característiques addicionals i els ajustos realitzats han contribuït positivament en la capacitat dels models per a comprendre i analitzar el text en diferents idiomes.

4 Altres codificacions

Ara provarem amb altres tipus de codificacions de entitats i veurem si el rendiment varia. Com que abans ja hem realitzat un grid_search, utilitzarem la millor combinació de característiques obtinguda anteriorment. I com que a més hem demostrat que la nostra features_getter obté uns millors resultats que la funció features_getter per defecte, utilitzarem la nostra pròpia.

4.1 Codificació IO

La codificació IO és inclús més senzilla que la BIO i simplement indica si una paraula forma part d'una entitat amb una 'I', i del contrari fica una 'O'.

Així doncs, hem creat una funció per passar de la codificació BIO a IO. Tant per l'idioma espanyol com per el neerlandès, hem entrenat un model amb aquesta nova codificació i posteriorment hem analitzat els resultats seguint els mateixos passos realitzats en la secció Model per predir entitats BIO.

4.1.1 Resultats

Idioma	Precisió	Recuperació	Puntuació F1
Espanyol	0.785	0.762	0.773
Neerlandès	0.746	0.685	0.714

Taula 2: Resultats codificació IO

Com podem veure els resultats són similars per el cas de l'espanyol i una mica pitjors per el cas de neerlandès.

Una possible explicació per a aquesta discrepància podria ser que, com que l'idioma neerlandès té característiques lingüístiques diferents de l'espanyol, els models podrien tenir més dificultats per a identificar les entitats en textos en neerlandès. Per exemple, les particularitats gramaticals, com la conjugació dels verbs o la formació de plurals, podrien afectar la capacitat del model per a generalitzar i reconèixer les entitats de manera precisa.

4.2 Codificació BIOW

La codificació BIOW és una extensió de la codificació BIO. En aquesta codificació, a més dels marcadors B (principi de l'entitat) i I (dins de l'entitat), s'afegeix un marcatge W per indicar que és una entitat amb una única paraula.

Hem creat una funció per passar de la codificació BIO a IO. Tant per l'idioma espanyol com per el neerlandès, hem entrenat un model amb aquesta nova codificació i posteriorment hem analitzat els resultats seguint els mateixos passos realitzats en la secció Model per predir entitats BIO.

Taula 3: Resultats del sistema de reconeixement d'entitats utilitzant la codificació BIOW per a l'idioma Espanyol i Neerlandès.

Idioma	Precisió	Recuperació	Puntuació F1
Espanyol	0.785	0.762	0.773
Neerlandès	0.746	0.685	0.714

4.2.1 Resultats

Els resultats obtinguts amb la codificació BIOW mostren similituds amb els resultats obtinguts amb les codificacions anteriors per a l'idioma Espanyol i Neerlandès.

5 Execució del Model amb Textos Reals

Per avaluar la capacitat del nostre model de reconèixer entitats en textos reals, hem implementat un procés de prova amb textos en espanyol i neerlandès. Els textos seleccionats per a aquestes proves són representatius de la variabilitat lingüística i inclouen diverses formes gramaticals i tipologies de text que esperem trobar en aplicacions del món real. Cal remarcar que hem traduït l'arxiu del text en espanyol a neerlandès, per tal de veure si reconeixia o no les mateixes entitats, ja que fixant-nos en els índexs dels resultats després d'aplicar la funció `extract_entities`, podríem veure si les entitats que s'han predit com a tal en el model en espanyol, també s'han predit de la mateixa manera en el model neerlandès, tot i que en alguns casos els índexs no coincidiran, ja que al traduïr l'idioma molt probablement s'hagin afegit o tret paraules, la qual cosa canviarà el nombre total de paraules del text i, conseqüentment, el número dels índex inicial i final de cada entitat en comparació a l'altre arxiu de text.

5.1 Preparació dels Dades

Els textos utilitzats per a les proves es llegeixen de la carpeta `dades_reals`, que conté els fitxers `esp.txt` i `ned.txt` per a espanyol i neerlandès, respectivament. El procés comença amb la lectura i tokenització dels textos utilitzant la funció `read_and_tokenize`, que converteix el text en llistes de tokens.

5.2 Etiquetatge de Text

Una vegada tokenitzats els textos, creem instàncies del model CRFTagger configurades amb la funció `_get_features` personalitzada per a cada idioma. El model per a l'espanyol està configurat amb l'objecte `FeatureGetter` que utilitza l'etiquetador POS `ct_POS_esp`, i similarment, el model per al neerlandès utilitza `ct_POS_ned`.

5.3 Reconeixement d'Entitats

Després del POS tagging, apliquem funcions per a extreure entitats del text etiquetat. Aquesta etapa és vital perquè avalua la capacitat del model de identificar correctament entitats com a noms de persones, localitzacions, organitzacions, etc. Com que ara no tenim com avaluar el rendiment del model, ens limitarem a fixar-nos en quines entitats han estat reconegudes, si s'han classificat bé, i si

els dos models han reconegut un nombre similar d'entitats, ja que com hem comentat anteriorment, són textos breus i un és la traducció de l'altre.

5.4 Resultats Obtinguts

En el text en espanyol, el model ha identificat un nombre significatiu d'entitats de diverses categories, incloent múltiples localitzacions i persones, així com organitzacions i algunes etiquetes MISC. Hem observat que ciutats com "Europa", "París" i "Francia" han estat correctament identificades com a LOC. Tot i així, algunes etiquetes podrien millorar, com l'etiquetatge de "Torre Eiffel" que ha estat reconeguda com PER en lloc de LOC. Això podria indicar una confusió del model entre noms propis de persones i els de monuments o localitzacions. L'identificació de organitzacions i altres entitats generalment ha funcionat bé, encara que hi ha estat algunes inconsistències, com la classificació de "Revolución Francesa" que ha estat parcialment etiquetada com ORG quan seria més apropiat classificar-la com a MISC. Per la qual cosa ens adonem que encara hi ha marge de millora alhora de perfeccionar el model.
Pel que fa al neerlandès, el model ha mostrat una eficàcia similar, amb correctes identificacions de localitzacions com "Europa", "Parijs", i "Frankrijk". No obstant això, de nou, "Eiffeltoren" ha estat etiquetat com MISC, suggerint una tendència del model a confondre la categoria de certes entitats famoses.
A més, la identificació de persones i organitzacions ha estat generalment precisa, amb algunes excepcions menors que podrien requerir una revisió del conjunt de dades d'entrenament o ajustaments en les característiques utilitzades pel model.
Tot i així, algunes categories com MISC semblen menys freqüents comparades amb les altres, el que podria indicar una àrea d'optimització per al model, ja sigui en l'entrenament o en l'ajustament de les característiques utilitzades per a la classificació d'entitats.
Els resultats indiquen que, mentre que els models poden reconèixer amb èxit moltes entitats correctament, hi ha àrees de millora en la identificació de la categoria correcta d'algunes entitats, particularment aquelles que poden pertànyer a múltiples categories. Tot i així, els models han demostrat ser efectius en la identificació de diverses classes d'entitats dins dels textos, la qual cosa valida la seva utilitat en aplicacions de processament de llenguatge natural que requereixen NER. Finalment, però, seria essencial revisar els casos on el model ha fallat o ha mostrat ambigüitats en l'etiquetatge per millorar la seva precisió si el volguéssim perfeccionar en un futur.