Université Paris Descartes

Projet pluridisciplinaire-MLSD

Biomedical Named Entity Recognition

Team:

ABIDAR Bouchra

BOUSSEBAINE Mustapha

LARBI Abderrahmane

Teacher:

AFFELDT Severine

2019-2020

Contents

1

Introduction

La reconnaissance d'entités nommées trouve leurs origines dans le domaine de la linguistique. Alors considérées comme sous-tâches de l'extraction d'information, ces disciplines ont rapidement attiré l'attention de différents domaines scientifiques tels que la biologie et la biomédecine. Avec le volume important des connaissances médicales numérisées à large échelle, retrouver automatiquement une information de haute précision est devenu un défi.

La reconnaissance biomédicale des entités nommées (Bio-NER) est une tâche fondamentale dans la gestion des termes textuels biomédicaux importants tels que les maladies, les gènes, les produits chimiques.

L'ensemble des études présente dans ce document, s'intéressent principalement à expérimenter les approches algorithmiques profonds en se basant sur les entités présentes sur différents corpus biomédicales.

Ce manuscrit est organisé selon quatre grandes parties. Nous commencerons par définir l'ensemble des algorithmes utilisés, ensuite les différents corpus expérimentés. Nous exploiterons ainsi les résultats obtenus sous forme d'une étude comparative. Puis nous présenterons un modèle que nous avons entraîné sur un corpus regroupant trois différentes thématiques. Ces dernières résumant l'ensemble des travaux réalisés jusqu'à présent. Néanmoins, nous finirons par proposer quelques techniques trouvées dans la littérature du NER pour améliorer la précision de la détection des entités nommées Biomédical.

2

Les algorithmes utilisés

2.1 CRF

2.1.1 CRF théoriquement

Les conditional Random Fields (ou CRF) sont des modèles Markoviens conditionnels qui ont été proposés pour remédier à certains défauts des modèles Markoviens plus classiques. Ils peuvent être appliqués à des données séquentielles, des données de type arbre et de façon plus générale à tout type de données structurées.

Les CRFs prennent comme entrés des données séquentielle avec une prise en compte de contexte précédent lors des prédictions sur un point de données. Nous modélisons mathématiquement cet aspect par *"features de connaissances"* définit comme suite :

$$f(X, i, l_{i-1}, l_i) \tag{1}$$

Le but de *features de connaissances* est de caractériser la séquence. Elle se base sur l'étiquette du mot précédent et du mot actuel. La construction du champ conditionnel est faite en attribuant à chaque *features de connaissances* un ensemble de poids (λ) que l'algorithme va apprendre. La fonction suivante explique la distribution de probabilité pour les champs aléatoires conditionnels CRF [1].

$$P(y, X, \lambda) = \frac{1}{Z(X)} \exp^{\sum_{i=1}^{n} \sum_j \lambda_j f_i(X, i, yi-1, yi)} \tag{2}$$

et :

$$Z(X) = \sum_{y' \in y} \sum_{i=1}^{n} \sum_j \lambda_j f_i(X, i, y'i - 1, y'i) \tag{3}$$

Les champs aléatoires conditionnels définissent les fonctions d'entité nécessaires, en initialisant les poids à des valeurs aléatoires, puis ils utilisent la descente de gradient de manière itérative jusqu'à la convergence des λ. L'équation de la mise à jour finale de descente de gradient pour

CRF est défini par l'équation (2) [1]:

$$\lambda = \lambda + a[\sum_{k=1}^{m} F_j(y^k, x^k) + \sum_{k=1}^{m} P(y|x^k, \lambda)F_j(y, x^k)] \quad (4)$$

Les CRF sont similaires à la régression logistique, car ils utilisent la distribution de probabilité conditionnelle. La différence est que l'algorithme CRF utilise "features de connaissances" comme entrées séquentielles.
Pour résumer, nous présenterons une synthèse d'utilisation un CRF.

- Il faut fournir des exemples(x, y).
- Il faut redéfinir sous forme de features ses connaissances.
- Il faut lancer l'apprentissage des K poids λ_k :
  1. En maximisant la log-vraisemblance
  2. Méthode : descente de gradient
- Une fois le modèle fixé, il permet de trouver l'annotation y qui maximise p(y|x) pour tout nouveau x.

2.1.2 CRF techniquement

Pour tester cette approche, nous avons expérimenté deux packages : *sklearn_crfsuite* et *pycrfsuit*. L'objectif est de pouvoir comparer les scores de performances des modelés.
Pour rajouter des informations aux différents corpus, nous avons utilisé le package *NLTK* pour avoir les entités *POS*. Les différents expérimentations de CRF sur les corpus de thématiques différents (maladies, gênes et produits chimiques) existent dans le repot git: code.

2.2 BiLSTM-CNN-CRF

2.2.1 BiLSTM-CNN-CRF théoriquement

Les études [2] ont montré que CNN est une approche efficace pour extraire des informations morphologiques(comme le préfixe ou le suffixe d'un mot) à partir de caractères de mots(figure 1).

Pour de nombreuses tâches d'étiquetage séquentiel, il est avantageux d'avoir accès aux deux contextes : passés et futurs. Cependant, une solution élégante dont l'efficacité a été prouvée par des travaux antérieurs [3] est le LSTM bidirectionnel (BLSTM).

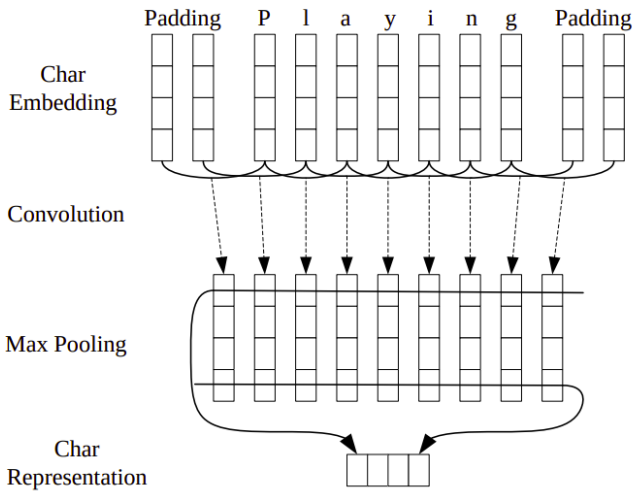


Figure 1: CNN pour l'extraction de représentations de mots au niveau des caractères.

L'idée de base est de présenter chaque séquence en avant et en arrière dans deux états cachés distincts afin de saisir les informations passées et futures, respectivement. Ensuite, les deux états cachés sont concaténés pour former le résultat final. La figure(2) illustre l'architecture d'un BiLSTM [4]. Vous pouvez consulter l'ensemble d'expérimentation réalisé par l'équipe dans le dépôt git code.
Le modèle BiLSTM-CNN-CRF utilise ces deux architectures :

- Encodeur CNN pour la représentation embeddings.
- LSTM bidirectionnel pour le codage au niveau du mot.
- Champs aléatoires conditionnels (CRF) pour le décodage de sortie.

La figure(3) présente l'architecture de modèle BiLSTM-CNN-CRF.

2.2.2 BiLSTM-CNN-CRF techniquement

Nous avons implémenté l'architecture BiLSTM CNN CRF en utilisant Keras Tensorflow. Nous avons utilisé le codage par CNN pour le codage par réseau neuronal à convolution pour la représentation des mots au niveau des caractères, le LSTM bidirectionnel pour le codage au niveau des mots et les champs aléatoires conditionnels (couche CRF) pour les décodages de sortie.

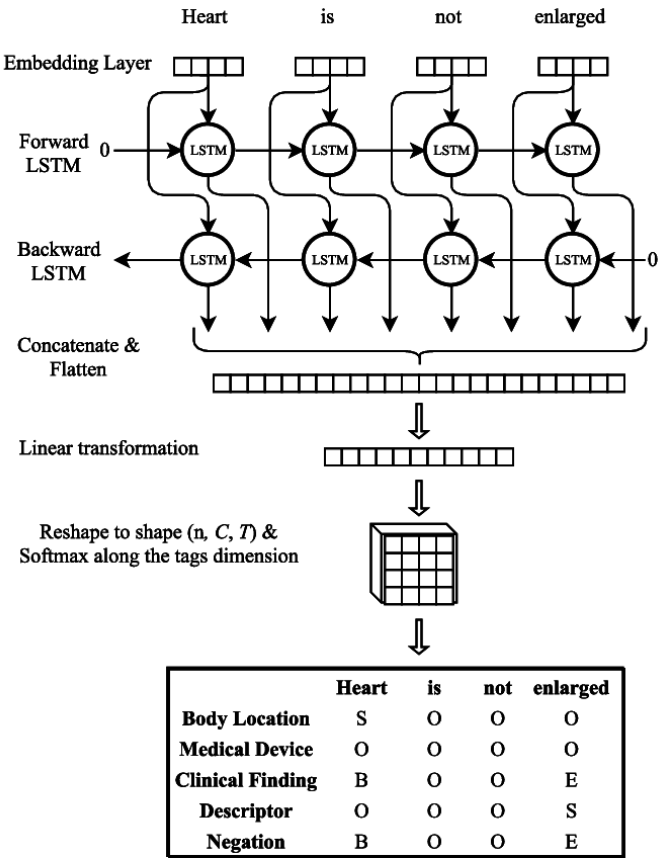


Figure 2: L'architecture de BLSTM

2.3 BERT

Bert a été présenté par Google. Ce modèle implémente plusieurs mécanismes notamment : *Transformer* et *Masked LM (MLM)*.

Transformer apprend les relations contextuelles entre les mots dans un texte[5]. Ce mécanisme contient deux parties : encodeur qui lit l'entrée de texte et décodeur pour prédire. Le mécanisme *Transformer* permet au modèle d'apprendre le contexte d'un mot en se basant sur l'ensemble de son environnement (gauche et droite du mot). Contrairement aux modèles directionnels, qui lisent l'entrée de texte séquentiellement (de gauche à droite ou de droite à gauche).

La deuxième stratégie qui rend Bert puissant est l'opération *Masked LM (MLM)*. Ce mécanisme permet de masquer 15% des mots de chaque séquence. Le modèle tente ensuite de prédire la valeur d'origine des mots masqués, sur la base du contexte(*Transformer*) fourni par les autres mots non masqués de la séquence (figure(4)).

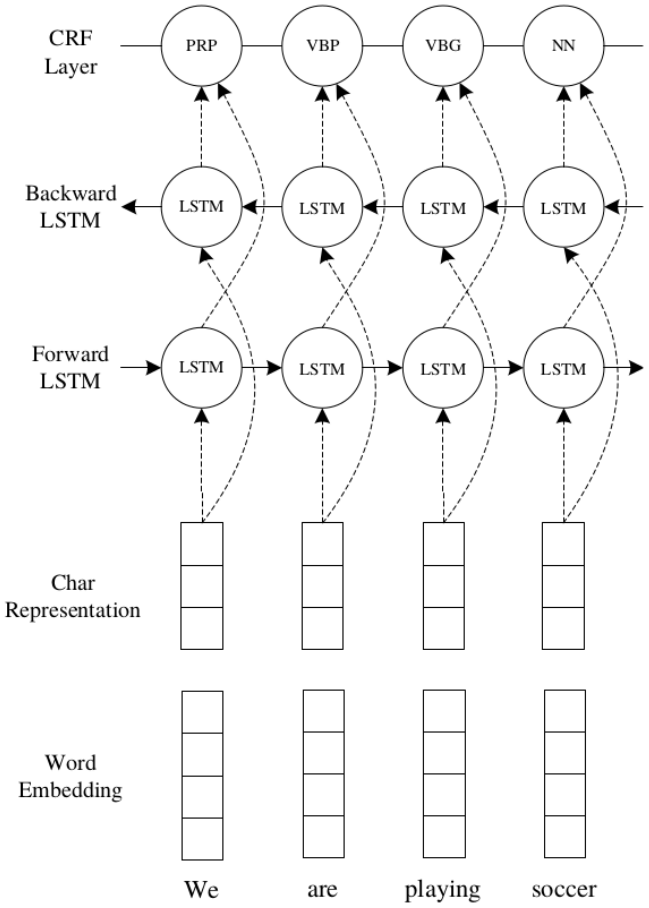


Figure 3: L'architecture de BLSTM-CNN-CRF[2]

2.3.1 Flair

Flair s'appuie sur une modélisation du langage neuronale au niveau des caractères pour apprendre des représentations contextuelles puissantes du langage humain à partir de grands corpus. La figure(5) présente l'architecture générale de Flair. Dans l'architecture de Flair, une phrase est vue en tant que séquence de caractères dans un modèle de langage de caractères bidirectionnel ML (Partie jaune de figure (5)) qui a été entraîné sur de très grands corpus de textes non étiquetés. À partir de ce LM, nous récupérons word a contextual embedding en enregistrant les états de la cellule du premier et du dernier caractère. Ce word embedding est ensuite passé dans un séquence BiLSTM-CRF vanille (bleu dans la figure 5).

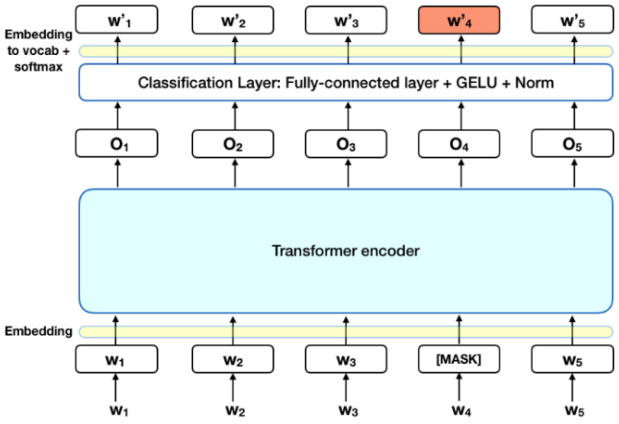


Figure 4: Mécaniseme du Masked LM[5]

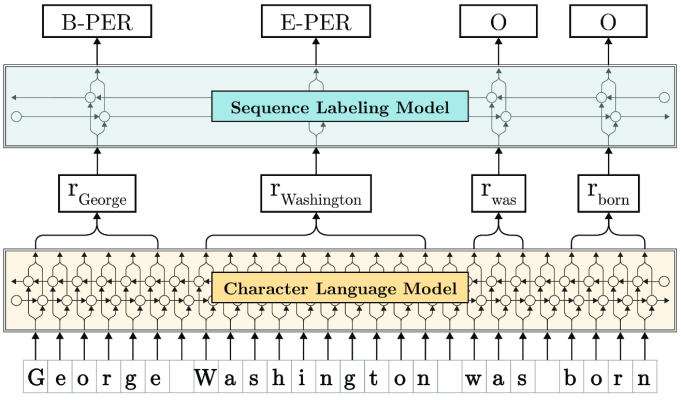


Figure 5: L'architecture de Flaire [6]

2.4 Spacy

Spacy est un outil pour la déction des NER. Il prend des décions sur une base d'une prédiction entrainée pendant l'apprentissage. La figure suivant(figure 6) [10] explique le fonctionnement de spacy. Le format des données d'entraînement pour le modéle *Spacy* est différent de ceux

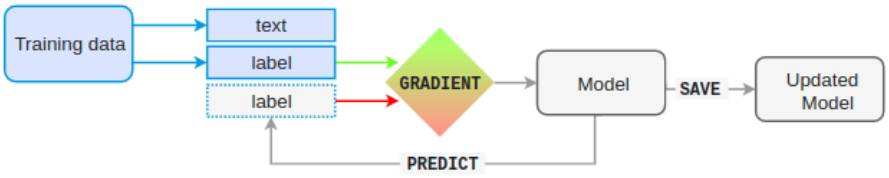


Figure 6: Le fonctionnement de spacy[10]

d'autres modéles. D'ou la necessité de procéder à une étape de pre-processing. Un fichier *helper.py* présent dans le repôt git code contient les différents fonctions permettant de convertir automatiquement les données sous le format adapté aus inputs Spacy.
Dans l'utilisation de Spacy, la seule métrique qui évalue le modèle est *Loss*(resultat de loss).
Les graphes suivants expliquent l'évolution de loss en fonction des epochs

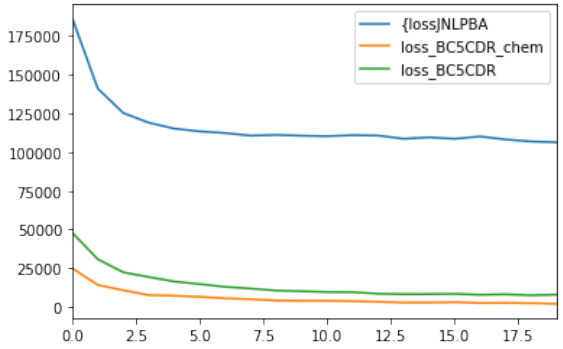


Figure 7: Loss spacy

3

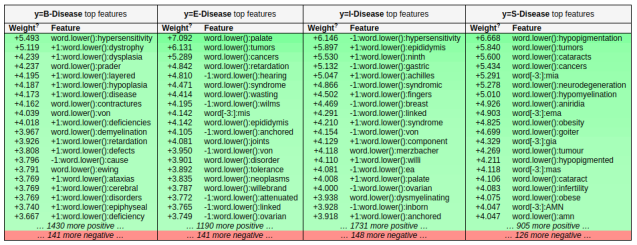
Les données

L'ensemble de données utilisées dans cette étude sont présentées dans le tableau ci-dessous(Table.1). Chaque corpus présente une thématique lié à un domaine bio-Médical notamment : les maladies, gênes, protéines et les produits chimiques.

DataSets					
corpus	Type d'Entité	Phrases	annotations	Taille	Publication
NCBI	Diseases	7639	6881	793	source
BC5CDR	Diseases	14228	12852	1500	source
JNLPBA	Genes	22562	35336	2404	source
BC5CDR-chem	Chemicals	14228	15935	1500	source

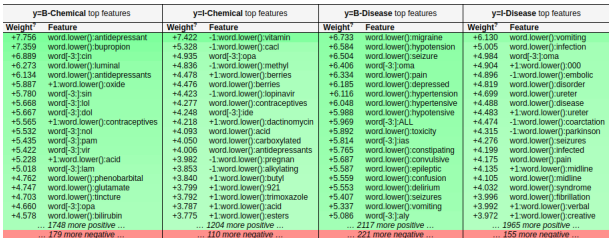
Table 1: Données d'entraînement

Les captures d'écrans suivantes présentent les top features dans les corpus (*NCBI, BC5CDR, BC5CDR-chemical et JNLPBA*) avec une estimation des poids associées(code)



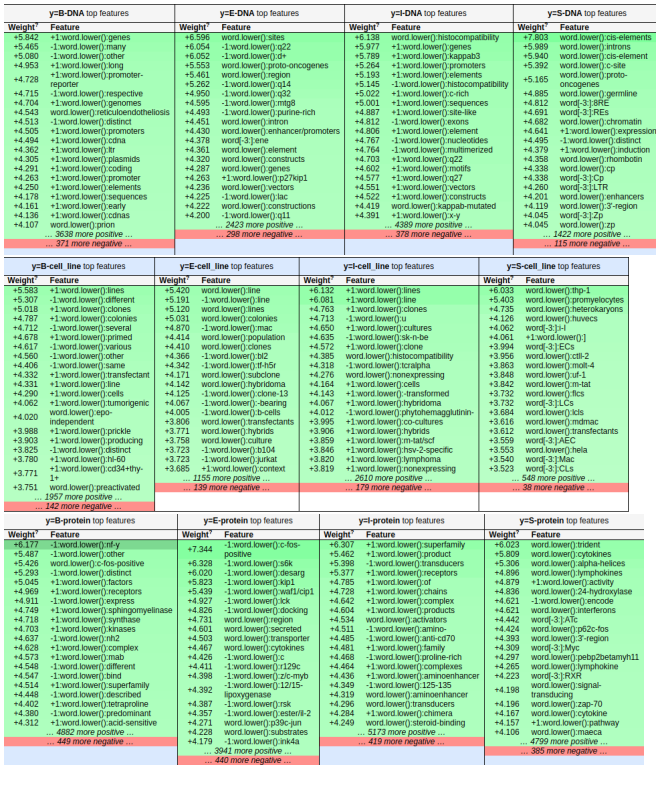
y=B-Disease top features		y=E-Disease top features		y=I-Disease top features		y=S-Disease top features	
Weight↑	Feature	Weight↑	Feature	Weight↑	Feature	Weight↑	Feature
+5.493	word.lower():hypersensitivity	+7.092	word.lower():palate	+6.146	-1:word.lower():hypersensitivity	+6.668	word.lower():hypopigmentation
+5.119	+1:word.lower():dystrophy	+6.131	word.lower():tumors	+5.897	+1:word.lower():epididymis	+5.840	word.lower():tumors
+4.239	+1:word.lower():dysplasia	+5.289	word.lower():cancers	+5.530	+1:word.lower():ninth	+5.600	word.lower():cataracts
+4.237	word.lower():prader	+4.842	word.lower():retardation	+5.132	-1:word.lower():gastric	+5.434	word.lower():cancers
+4.195	+1:word.lower():layered	+4.810	-1:word.lower():hearing	+5.047	+1:word.lower():achilles	+5.291	word:[-3]:mia
+4.187	+1:word.lower():hypoplasia	+4.471	word.lower():syndrome	+4.866	-1:word.lower():syndromic	+5.278	word.lower():neurodegeneration
+4.173	+1:word.lower():disease	+4.414	word.lower():wasting	+4.502	+1:word.lower():fingers	+5.010	word.lower():hypomyelination
+4.162	word.lower():contractures	+4.195	-1:word.lower():wilms	+4.469	-1:word.lower():breast	+4.926	word.lower():aninidia
+4.039	word.lower():von	+4.142	word:[-3]:mis	+4.291	-1:word.lower():linked	+4.903	word:[-3]:ema
+4.018	+1:word.lower():deficiencies	+4.142	word.lower():epididymis	+4.210	+1:word.lower():syndrome	+4.825	word.lower():obesity
+3.967	word.lower():demyelination	+4.105	-1:word.lower():anchored	+4.154	-1:word.lower():von	+4.699	word.lower():goiter
+3.926	+1:word.lower():retardation	+4.081	word.lower():joints	+4.129	+1:word.lower():component	+4.329	word:[-3]:gia
+3.808	+1:word.lower():defects	+3.950	-1:word.lower():von	+4.118	word.lower():merzbacher	+4.269	word.lower():tumour
+3.796	-1:word.lower():cause	+3.901	word.lower():disorder	+4.110	+1:word.lower():willi	+4.211	word.lower():hypopigmented
+3.791	word.lower():ewing	+3.892	word.lower():tolerance	+4.081	-1:word.lower():ea	+4.118	word:[-3]:mas
+3.769	+1:word.lower():ataxias	+3.835	word.lower():neoplasms	+4.008	+1:word.lower():palate	+4.106	word.lower():cataract
+3.769	+1:word.lower():cerebral	+3.787	word.lower():wilebrand	+4.000	-1:word.lower():ovarian	+4.083	word.lower():infertility
+3.769	+1:word.lower():disorders	+3.772	-1:word.lower():attenuated	+3.938	word.lower():dysmyelinating	+4.075	word.lower():obese
+3.740	+1:word.lower():epiphyseal	+3.765	-1:word.lower():linked	+3.928	-1:word.lower():inborn	+4.047	word:[-3]:AMN
+3.667	+1:word.lower():deficiency	+3.749	-1:word.lower():ovarian	+3.918	+1:word.lower():anchored	+4.047	word.lower():amn
	... 1430 more positive ...		*... 1190 more positive ...*		*... 1731 more positive ...*		*... 905 more positive ...*
	... 141 more negative ...		*... 141 more negative ...*		*... 148 more negative ...*		*... 126 more negative ...*

Figure 8: *NCBI top features*



y=B-Chemical top features		y=I-Chemical top features		y=B-Disease top features		y=I-Disease top features	
Weight↑	Feature	Weight↑	Feature	Weight↑	Feature	Weight↑	Feature
+7.756	word.lower():antidepressant	+7.422	-1:word.lower():vitamin	+6.733	word.lower():migraine	+6.130	word.lower():migraine
+7.359	word.lower():bupropion	+5.328	-1:word.lower():cacl	+6.584	word.lower():hypotension	+6.130	word.lower():vomiting
+6.889	word:[-3]:cin	+4.935	word:[-3]:opa	+6.504	word.lower():seizure	+5.005	word.lower():infection
+6.273	word.lower():luminal	+4.836	-1:word.lower():methyl	+6.406	word:[-3]:oma	+4.984	+1:word.lower():000
+6.134	word.lower():antidepressants	+4.478	+1:word.lower():berries	+6.334	word.lower():pain	+4.896	-1:word.lower():embolic
+5.887	+1:word.lower():oxide	+4.476	word.lower():berries	+6.185	word.lower():depressed	+4.819	word.lower():disorder
+5.780	word:[-3]:sin	+4.423	-1:word.lower():lopinavir	+6.116	word.lower():hypertension	+4.699	word.lower():ureter
+5.668	word:[-3]:iol	+4.277	word.lower():contraceptives	+6.048	word.lower():hypertensive	+4.488	word.lower():disease
+5.667	word:[-3]:dol	+4.248	word:[-3]:ide	+5.988	word.lower():hypotensive	+4.483	+1:word.lower():ureter
+5.565	+1:word.lower():contraceptives	+4.218	+1:word.lower():dactinomycin	+5.969	word:[-3]:ALL	+4.474	-1:word.lower():coarctation
+5.532	word:[-3]:nol	+4.093	word.lower():acid	+5.892	word.lower():toxicity	+4.315	-1:word.lower():parkinson
+5.435	word:[-3]:pam	+4.050	word.lower():carboxylated	+5.814	word:[-3]:ias	+4.276	word.lower():seizures
+5.422	word:[-3]:vir	+4.006	word.lower():antidepressants	+5.765	word.lower():constipating	+4.199	word.lower():infected
+5.228	+1:word.lower():acid	+3.982	-1:word.lower():pregnan	+5.687	word.lower():convulsive	+4.175	word.lower():pain
+5.018	word:[-3]:lam	+3.853	-1:word.lower():alkylating	+5.587	word.lower():epileptic	+4.135	+1:word.lower():midline
+4.762	word.lower():phenobarbital	+3.840	+1:word.lower():butyl	+5.559	word.lower():confusion	+4.105	word.lower():midline
+4.747	word.lower():glutamate	+3.799	+1:word.lower():921	+5.553	word.lower():delirium	+4.032	word.lower():syndrome
+4.703	word.lower():tincture	+3.792	+1:word.lower():trimoxazole	+5.407	word.lower():seizures	+3.996	word.lower():fibrillation
+4.660	word:[-3]:opa	+3.787	+1:word.lower():acid	+5.327	word.lower():vomiting	+3.992	+1:word.lower():verbal
+4.578	word.lower():bilirubin	+3.775	+1:word.lower():esters	+5.086	word:[-3]:aly	+3.972	+1:word.lower():creative
	... 1748 more positive ...		*... 1204 more positive ...*		*... 2117 more positive ...*		*... 1965 more positive ...*
	... 179 more negative ...		*... 110 more negative ...*		*... 221 more negative ...*		*... 155 more negative ...*

Figure 10: *BC5CDR chemical top features*



y=B-DNA top features		y=E-DNA top features		y=I-DNA top features		y=S-DNA top features	
Weight↑	Feature	Weight↑	Feature	Weight↑	Feature	Weight↑	Feature
+5.542	+1:word.lower():genes	+6.596	word.lower():sites	+6.138	word.lower():histocompatibility	+7.803	word.lower():cis-elements
+5.465	-1:word.lower():many	+6.054	-1:word.lower():q22	+5.977	+1:word.lower():genes	+5.989	word.lower():introns
+5.080	-1:word.lower():other	+6.052	-1:word.lower():d+	+5.789	+1:word.lower():kappab3	+5.940	word.lower():cis-element
+4.953	+1:word.lower():long	+5.553	word.lower():proto-oncogenes	+5.264	+1:word.lower():promoters	+5.392	word.lower():c-site
+4.728	+1:word.lower():promoter-reporter	+5.461	word.lower():region	+5.193	+1:word.lower():elements	+5.165	word.lower():proto-oncogenes
+4.715	-1:word.lower():respective	+5.262	-1:word.lower():q14	+5.145	-1:word.lower():histocompatibility	+4.885	word.lower():germline
+4.704	+1:word.lower():genomes	+4.950	-1:word.lower():q32	+5.022	+1:word.lower():c-rich	+4.682	word.lower():germline
+4.543	word.lower():reticuloendotheliosis	+4.595	-1:word.lower():mtg8	+5.001	+1:word.lower():sequences	+4.812	word:[-3]:8RE
+4.513	-1:word.lower():distinct	+4.493	-1:word.lower():purine-rich	+4.887	+1:word.lower():site-like	+4.691	word:[-3]:REs
+4.505	+1:word.lower():promoters	+4.451	word.lower():intron	+4.812	-1:word.lower():exons	+4.682	word.lower():chromatin
+4.494	+1:word.lower():cdna	+4.430	word.lower():enhancer/promoters	+4.806	+1:word.lower():element	+4.641	+1:word.lower():expression
+4.362	+1:word.lower():ltr	+4.378	word:[-3]:ene	+4.767	-1:word.lower():nucleotides	+4.495	-1:word.lower():distinct
+4.305	+1:word.lower():plasmids	+4.361	word.lower():element	+4.764	-1:word.lower():multimerized	+4.379	+1:word.lower():induction
+4.291	+1:word.lower():coding	+4.320	word.lower():constructs	+4.703	+1:word.lower():q22	+4.358	word.lower():rhombotin
+4.263	+1:word.lower():promoter	+4.287	word.lower():genes	+4.602	+1:word.lower():motifs	+4.338	word.lower():cp
+4.250	+1:word.lower():elements	+4.263	+1:word.lower():p27kip1	+4.577	+1:word.lower():q27	+4.338	word:[-3]:Cp
+4.178	+1:word.lower():sequences	+4.236	word.lower():vectors	+4.551	+1:word.lower():vectors	+4.260	word:[-3]:LTR
+4.161	+1:word.lower():early	+4.225	-1:word.lower():lac	+4.522	+1:word.lower():constructs	+4.201	word.lower():enhancers
+4.136	+1:word.lower():cdnas	+4.222	word.lower():constructions	+4.419	word.lower():kappab-mutated	+4.119	word.lower():3'-region
+4.107	word.lower():prion	+4.200	-1:word.lower():q11	+4.391	+1:word.lower():x-y	+4.045	word:[-3]:Zp
	... 3638 more positive ...		*... 2423 more positive ...*		*... 4389 more positive ...*		*... 1422 more positive ...*
	... 371 more negative ...		*... 298 more negative ...*		*... 378 more negative ...*		*... 115 more negative ...*

y=B-cell_line top features		y=E-cell_line top features		y=I-cell_line top features		y=S-cell_line top features	
Weight↑	Feature	Weight↑	Feature	Weight↑	Feature	Weight↑	Feature
+5.583	+1:word.lower():lines	+5.420	word.lower():line	+6.132	+1:word.lower():lines	+6.033	word.lower():thp-1
+5.307	-1:word.lower():different	+5.191	-1:word.lower():line	+6.081	+1:word.lower():line	+5.403	word.lower():promyelocytes
+5.018	+1:word.lower():clones	+5.120	word.lower():lines	+4.763	+1:word.lower():clones	+4.735	word.lower():heterokaryons
+4.797	+1:word.lower():colonies	+5.031	word.lower():colonies	+4.713	-1:word.lower():u	+4.126	word.lower():huvecs
+4.712	-1:word.lower():several	+4.870	-1:word.lower():mac	+4.650	+1:word.lower():cultures	+4.062	word:[-3]:i-l
+4.678	+1:word.lower():primed	+4.414	word.lower():population	+4.635	-1:word.lower():sk-n-be	+4.061	+1:word.lower():]
+4.617	-1:word.lower():various	+4.410	word.lower():clones	+4.572	+1:word.lower():clone	+3.994	word:[-3]:ECs
+4.560	-1:word.lower():other	+4.366	-1:word.lower():b2	+4.385	word.lower():histocompatibility	+3.956	word.lower():ctll-2
+4.406	-1:word.lower():same	+4.342	-1:word.lower():tf-h5r	+4.318	-1:word.lower():tcralpha	+3.863	word.lower():molt-4
+4.332	+1:word.lower():transfectant	+4.171	word.lower():subclone	+4.276	word.lower():nonexpressing	+3.848	word.lower():ut-1
+4.331	+1:word.lower():line	+4.142	word.lower():hybridoma	+4.164	+1:word.lower():cells	+3.842	word.lower():m-tat
+4.290	+1:word.lower():cells	+4.125	-1:word.lower():clone-13	+4.143	+1:word.lower():-transformed	+3.732	word.lower():flcs
+4.062	+1:word.lower():tumorigenic	+4.067	-1:word.lower():-bearing	+4.067	+1:word.lower():hybridoma	+3.732	word:[-3]:LCs
+4.020	word.lower():epo-independent	+4.005	-1:word.lower():b-cells	+4.012	-1:word.lower():phytohemagglutinin-	+3.684	word.lower():kcls
+3.988	+1:word.lower():prickle	+3.806	word.lower():transfectants	+3.995	+1:word.lower():co-cultures	+3.616	word.lower():mdmac
+3.903	+1:word.lower():producing	+3.771	word.lower():hybrids	+3.906	+1:word.lower():hybrids	+3.612	word.lower():transfectants
+3.825	-1:word.lower():distinct	+3.758	word.lower():mare	+3.859	+1:word.lower():m-tatfast	+3.559	word:[-3]:AEC
+3.780	+1:word.lower():hi-60	+3.723	-1:word.lower():b104	+3.846	+1:word.lower():hsv-2-specific	+3.553	word.lower():hela
+3.771	+1:word.lower():cd34+thy-1+	+3.723	-1:word.lower():jurkat	+3.820	+1:word.lower():lymphoma	+3.540	word:[-3]:Mac
+3.751	word.lower():preactivated	+3.685	-1:word.lower():primed	+3.819	+1:word.lower():nonexpressing	+3.523	word:[-3]:CLs
	... 1957 more positive ...		*... 1155 more positive ...*		*... 2610 more positive ...*		*... 548 more positive ...*
	... 142 more negative ...		*... 139 more negative ...*		*... 179 more negative ...*		*... 38 more negative ...*

y=B-protein top features		y=E-protein top features		y=I-protein top features		y=S-protein top features	
Weight↑	Feature	Weight↑	Feature	Weight↑	Feature	Weight↑	Feature
+6.177	-1:word.lower():nf-y	+7.344	-1:word.lower():c-fos-positive	+6.307	+1:word.lower():superfamily	+6.023	word.lower():trident
+5.487	-1:word.lower():other			+5.462	+1:word.lower():product	+5.809	word.lower():cytokines
+5.426	word.lower():c-fos-positive	+6.328	-1:word.lower():s6k	+5.398	-1:word.lower():transducers	+5.306	word.lower():alpha-helices
+5.293	-1:word.lower():distinct	+6.020	-1:word.lower():desarg	+5.377	+1:word.lower():receptors	+4.896	word.lower():lymphokines
+5.045	+1:word.lower():factors	+5.823	-1:word.lower():kip1	+4.785	+1:word.lower():of	+4.879	+1:word.lower():activity
+4.969	+1:word.lower():receptors	+5.439	-1:word.lower():waf1/cip1	+4.728	+1:word.lower():chains	+4.836	word.lower():24-hydroxylase
+4.911	-1:word.lower():express	+4.927	-1:word.lower():ick	+4.642	+1:word.lower():complex	+4.621	-1:word.lower():encode
+4.749	+1:word.lower():synmyelinase	+4.826	-1:word.lower():docking	+4.604	+1:word.lower():products	+4.621	word.lower():interferons
+4.718	+1:word.lower():synthase	+4.731	word.lower():region	+4.534	word.lower():activators	+4.442	word:[-3]:ATc
+4.703	+1:word.lower():kinases	+4.601	word.lower():secreted	+4.511	-1:word.lower():amino-	+4.424	word.lower():p62c-fos
+4.637	-1:word.lower():nh2	+4.503	word.lower():transporter	+4.485	-1:word.lower():anti-cd70	+4.393	word.lower():3'-region
+4.628	+1:word.lower():complex	+4.467	word.lower():cytokines	+4.481	+1:word.lower():family	+4.309	word:[-3]:Myc
+4.573	+1:word.lower():mab	+4.426	-1:word.lower():c	+4.468	-1:word.lower():proline-rich	+4.297	word.lower():pebp2betamyh11
+4.548	-1:word.lower():different	+4.411	-1:word.lower():r129c	+4.464	+1:word.lower():complexes	+4.265	word.lower():lymphokine
+4.547	-1:word.lower():bind	+4.398	-1:word.lower():z/c-myb	+4.436	+1:word.lower():aminoenhancer	+4.223	word:[-3]:RXR
+4.514	+1:word.lower():superfamily	+4.392	-1:word.lower():12/15-lipoxygenase	+4.349	-1:word.lower():125-135	+4.198	word.lower():signal-transducing
+4.448	-1:word.lower():described			+4.319	word.lower():aminoenhancer		
+4.402	+1:word.lower():tetraproline	+4.387	-1:word.lower():rsk	+4.296	word.lower():transducers	+4.196	word.lower():zap-70
+4.380	-1:word.lower():predominant	+4.357	-1:word.lower():ester/ll-2	+4.284	+1:word.lower():chimera	+4.167	word.lower():cytokine
+4.312	+1:word.lower():acid-sensitive	+4.271	word.lower():p39c-jun	+4.249	word.lower():steroid-binding	+4.157	+1:word.lower():pathway
	... 4882 more positive ...	+4.228	word.lower():substrates		*... 5173 more positive ...*	+4.106	word.lower():maeca
	... 449 more negative ...	+4.179	-1:word.lower():nik4a		*... 419 more negative ...*		*... 4799 more positive ...*
			... 3941 more positive ...				*... 385 more negative ...*
			... 440 more negative ...				

Figure 11: *Genes top features*

y=B-RNA top features		y=E-RNA top features		y=I-RNA top features		y=S-RNA top features	
Weight↑	Feature	Weight↑	Feature	Weight↑	Feature	Weight↑	Feature
+7.600	+1:word.lower():mma	+5.971	-1:word.lower():transcript	+7.183	+1:word.lower():mmas	+6.158	word:[-3]:RNA
+5.768	+1:word.lower():mmas	+5.707	word:[-3]:RNA	+6.108	+1:word.lower():transcripts	+4.522	word.lower():mmas
+5.731	+1:word.lower():transcripts	+5.130	-1:word.lower():mma	+6.025	+1:word.lower():mma	+4.298	word:[-3]:LF1
+4.709	-1:word.lower():express	+4.808	-1:word.lower():exon	+5.701	-1:word.lower():poly	+4.125	word:[-3]:ERs
+4.608	+1:word.lower():nas	+4.724	word:[-3]:NAs	+5.336	+1:word.lower():messenger	+4.125	word.lower():ebers
+4.608	word.lower():cepsilon	+4.679	-1:word.lower():ribonucleic	+4.950	+1:word.lower():alpha/mad-3	+3.862	word.lower():tar
+4.510	+1:word.lower():transcript	+4.675	word.lower():mma	+4.729	-1:word.lower():lamin	+3.835	word:[-3]:TAR
+4.312	+1:word.lower():polyadenylation	+4.615	-1:word.lower():g0s	+4.524	-1:word.lower():pafr	+3.692	word:[-3]:NAs
+4.212	word.lower():g0s	+4.002	word.lower():rna	+4.345	-1:word.lower():spermine	+3.540	word.lower():il-2rmma
+4.212	word:[-3]:G0S	+3.983	-1:word.lower():polyadenylation	+4.337	+1:word.lower():transcript	+3.449	word.lower():mrna
+4.167	+1:word.lower():myb	+3.956	-1:word.lower():secreted	+4.239	-1:word.lower():185+/-38	+3.435	-1:word.lower():express
+4.078	+1:word.lower():species	+3.779	-1:word.lower():re	+4.198	-1:word.lower():histocytic	+3.230	word:[-3]:15T
+3.922	+1:word.lower():ma	+3.769	word.lower():mmas	+4.179	+1:word.lower():er-beta	+3.230	word.lower():g2715t
+3.912	+1:word.lower():forms	+3.586	-1:word.lower():-induced	+3.991	+1:word.lower():er-beta	+3.096	-1:word.lower():induce
+3.782	-1:word.lower():affect	+3.561	word.lower():tnf-alpha-mrna	+3.842	word.lower():er-beta	+3.009	word.lower():cytokine-mrna
+3.775	word.lower():secreted	+3.461	word.lower():transcripts	+3.813	-1:word.lower():er-alpha	+2.997	-1:word.lower():specific
+3.757	+1:word.lower():messenger	+3.458	word:[-3]:pts	+3.805	word.lower():polyadenylylation	+2.994	-1:word.lower():inflammatory
+3.744	+1:word.lower():histocytic	+3.426	word.lower():truncation	+3.804	+1:word.lower():rna	+2.841	word:[-3]:mu
+3.670	+1:word.lower():delta-aminolevulinate	+3.376	-1:word.lower():ma	+3.723	+1:word.lower():ribonucleic	+2.841	word.lower():mu
+3.622	-1:word.lower():expression	+3.130	word.lower():mas	+3.680	+1:word.lower():cdlia	+2.716	word:[-3]:nsA
	... 680 more positive ...		*... 308 more positive ...*		*... 65 more positive ...*		*... 125 more positive ...*
	... 69 more negative ...		*... 65 more negative ...*		*... 63 more negative ...*		*... 28 more negative ...*

y=B-cell_type top features		y=E-cell_type top features		y=I-cell_type top features		y=S-cell_type top features	
Weight↑	Feature	Weight↑	Feature	Weight↑	Feature	Weight↑	Feature
+4.831	+1:word.lower():rbc	+4.376	word.lower():fraction	+5.528	+1:word.lower():lineage	+6.224	word:[-3]:ECs
+4.756	+1:word.lower():lineage	+4.300	word:[-3]:MCs	+4.935	-1:word.lower():glial	+5.624	word:[-3]:PCs
+4.636	+1:word.lower():fraction	+4.286	+1:word.lower():activity	+4.465	+1:word.lower():cd8+	+5.513	word:[-3]:MCs
+4.619	-1:word.lower():different	+4.268	-1:word.lower():erythroblastic	+4.386	+1:word.lower():cells	+4.839	word:[-3]:MNL
+4.592	word.lower():tartrate-resistant	+4.206	word.lower():progenitors	+4.378	-1:word.lower():cd8-	+4.781	word.lower():t-cells
		+3.803	word.lower():leucocytes	+4.304	+1:word.lower():mcd14-expressing	+4.613	word.lower():non-t-cells
+4.404	+1:word.lower():cells	+3.719	-1:word.lower():freshly	+4.265	+1:word.lower():types	+4.421	-1:word.lower():possible
+4.224	word.lower():adipocytes	+3.705	word.lower():population	+4.201	+1:word.lower():lineages	+4.179	word:[-3]:Th1
+4.218	-1:word.lower():other	+3.683	-1:word.lower():nf-kappab-mediated	+4.193	+1:word.lower():lymphocytes	+4.170	word:[-3]:yte
+4.212	+1:word.lower():platelets			+4.116	+1:word.lower():lesser	+4.150	word.lower():neuroblastoma
+4.159	+1:word.lower():sideroblasts	+3.591	-1:word.lower():t-b-1	+4.068	-1:word.lower():nk-	+4.079	word.lower():b-cells
+4.151	word.lower():professional	+3.542	-1:word.lower():adipocytes	+4.046	-1:word.lower():memory	+4.055	word:[-3]:ils
+4.040	-1:word.lower():various	+3.535	-1:word.lower():pmc	+4.042	+1:word.lower():mastocytic/basophilic	+4.004	word.lower():gcrbeta-positive
+4.015	+1:word.lower():progenitor	+3.527	word.lower():progenitor	+4.042	+1:word.lower():basosophilic		
+3.984	+1:word.lower():lymphocytes	+3.513	word.lower():lineage	+3.976	-1:word.lower():tartrate-resistant	+3.967	word.lower():pbmcs
+3.954	-1:word.lower():such	+3.496	-1:word.lower():hematopoietic	+3.949	-1:word.lower():cd45ro+	+3.936	word.lower():astrocytes
+3.923	+1:word.lower():dc	+3.485	word.lower():ecs	+3.913	-1:word.lower():helper	+3.853	word.lower():pbmc
+3.908	word.lower():intrafollicular	+3.430	+1:word.lower():identification	+3.775	word.lower():megakaryocyte	+3.853	word:[-3]:BMC
+3.900	+1:word.lower():peripheral	+3.320	+1:word.lower():differentiation	+3.768	-1:word.lower():cd45ra+	+3.850	word:[-3]:hil
+3.861	word.lower():oncogene	+3.282	word:[-3]:yte	+3.766	+1:word.lower():hla-dr-restricted	+3.843	word:[-3]:gas
+3.823	word.lower():antigen-presenting	+3.276	-1:word.lower():platelet-like			+3.820	word:[-3]:FDC
	... 1015 more positive ...		*... 1030 more positive ...*		*... 1657 more positive ...*		*... 638 more positive ...*
	... 105 more negative ...		*... 150 more negative ...*		*... 86 more negative ...*		*... 86 more negative ...*

4

Étude comparative

La table(2) résume les performances de détection d'entités de chaque algorithme.
En se basant sur la dispersion des classes, nous avons utilisé le *F1 Score* comme un indice de performance. Dans Les 4 corpus testés, FLair renvoie le meilleur score.

Table 2: La performances des modéles

Algorithmes		**F1 score(%)**
CRF	NCBI	77
	BC5CDR	75
	JNLPBA	70
	BC5CDR-chem	76
BiLSTM-CRF	NCBI	70
	BC5CDR	65
	JNLPBA	
	BC5CDR-chem	60
Flair	NCBI	88
	BC5CDR	88
	JNLPBA	81
	BC5CDR-chem	93

L'algorithmes *BiLSTM-CRF* a donné les faibles scores pour la détection des produits chimiques et les maladies sur le corpus BC5CDR. Nous proposons dans la section 6 **"Amélioration possibles"** des techniques trouvées dans la littérature pour augmenter la performance.
Suite à un problème au niveau des ressources, nous n'avons pas réussi à entraîner le modèle BiLSTM-CRF sur le corpus *JNLPBA*.

Une tentative a été testée en diminuant la taille des corpus. Cette tentative n'a pas abouté aux résultats attendus : F1 score égale à 25%
L'exemple d'éxecution des modèles sur des phrases différentes en temps réel est présenté dans la figure (12)

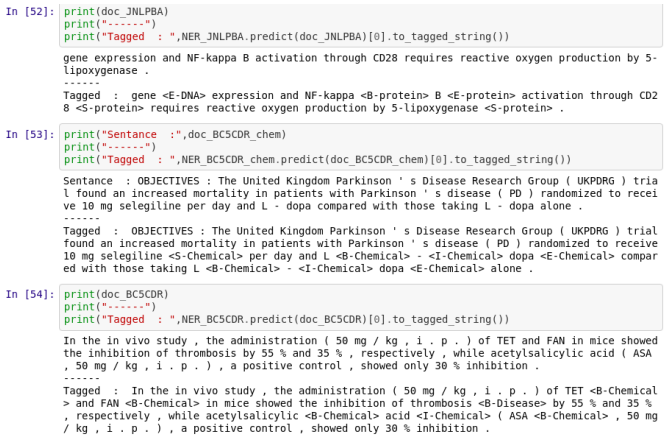


Figure 12: tester les modèles sur des phrases différentes

5

Modèle généralisé

Dans le but de pousser notre étude et expérimenter sur un corpus contenant les différentes thématiques abordées auparavant, nous avons décidé de créer un modèle entraîné sur un corpus regroupant les quatre thématiques. Le but est d'avoir un unique modèle capable de détecter les différents tags notamment : maladies, gènes et produits chimiques.
Le modèle Flair a été retenu pour cette tâche, suite a son résultat pertinent dans les corpus précédents ainsi l'intuitive de l'architecture de ce dernier.

EPOCH	LOSS	PRECISION	RECALL	F1
1	6.93	0.6386	0.4443	0.5240
10	3.22	0.7955	0.6725	0.7289
20	2.80	0.7664	0.7846	0.7754
50	2.54	0.7812	0.7929	0.7870
72	2.53	0.7820	0.7955	0.7887

Table 3: Résultats d'entraînement du modèle générale

Avec l'entraînement du modèle sur ce corpus généralisé, nous avons attient un score F1 de 79%. Un score qui reste intéressant en prenant en considération la distribution des classes non équilibrées ainsi que le nombre important de tagges différents présents dans le corpus construit.
Une autre manière de tester le modèle généralisé est d'essayer de tagger les features dans les mêmes phrases de figure(12). Ce que nous proposons dans la figure (13) code.

```
Sentance  :  gene expression and NF-kappa B activation through CD28 requires reactive oxygen pro
duction by 5-lipoxygenase .
------
Tagged  :  gene expression and NF-kappa <B-protein> B <E-protein> activation through CD28 <S-pro
tein> requires reactive oxygen production by 5-lipoxygenase .
--------
Sentance  :  In the in vivo study , the administration ( 50 mg / kg , i . p . ) of TET and FAN i
n mice showed the inhibition of thrombosis by 55 % and 35 % , respectively , while acetylsalicyl
ic acid ( ASA , 50 mg / kg , i . p . ) , a positive control , showed only 30 % inhibition .
------
Tagged  :  In the in vivo study , the administration ( 50 mg / kg , i . p . ) of TET <B-Chemical
> and FAN <B-Chemical> in mice showed the inhibition of thrombosis <B-Disease> by 55 % and 35 %
, respectively , while acetylsalicylic <B-Chemical> acid <I-Chemical> ( ASA <B-Chemical> , 50 mg
/ kg , i . p . ) , a positive control , showed only 30 % inhibition .
--------
Sentance  :  OBJECTIVES : The United Kingdom Parkinson ' s Disease Research Group ( UKPDRG ) tri
al found an increased mortality in patients with Parkinson ' s disease ( PD ) randomized to rece
ive 10 mg selegiline per day and L - dopa compared with those taking L - dopa alone .
------
Tagged  :  OBJECTIVES : The United Kingdom Parkinson <B-Disease> ' <I-Disease> s <I-Disease> Dis
ease <I-Disease> Research Group ( UKPDRG ) trial found an increased mortality in patients with P
arkinson <B-Disease> ' <I-Disease> s <I-Disease> disease <I-Disease> ( PD <B-Disease> ) randomiz
ed to receive 10 mg selegiline <B-Chemical> per day and L <B-Chemical> - <I-Chemical> dopa <I-Ch
emical> compared with those taking L <B-Chemical> - <I-Chemical> dopa <I-Chemical> alone .
```

Figure 13: Exemples des phrases tagger avec le modèle généralisé

6

Amélioration possibles

6.1 L'ajout des features

Nous pouvons toujours trouver dans la littérature des techniques pour améliorer le score de la performance des modèles. L'ensemble des travaux effectués dans l'article [7], montre une réussite remarquable d'incrémenter le score F1 en ajoutant des features pour améliorer les performances d'un CRF. Ce que nous pouvons faire aussi dans notre cas.
Une autre approche consiste à faire varier l'architecture de BiLSTM-CRF pour essayer d'augmenter le degré de précision de moddéle. Ce dépôt git présente un ensemble des architectures code

6.2 Changement d'architecture

L'une des architectures qui a donné des scores de performance remarquable environ 90% est BERT-BiLSTM-CRF-NER. Le lien suivant illustre les différents travaux présentés par les chercheurs code. La figure (14) présente l'architecture de ce modèle.

6.3 Apprentissage continuée

Une autre approche proposée pour augmenter la performance de la détection des NER consiste à créer un **Meta-Modèle** qui regroupe le modèle *Flair*, *Spacy*, *CRF* et *BLSTM*. L'idée derrière ce regroupement est de donner la possibilité à chaque modèle d'apprendre en permanence en fonction des résultats des autres modèles. Nous pouvons mettre en place cette approche en implémentant les étapes suivantes :

- Le modèle prend comme entrée une phrase à tagger.
- La phrase en question passe par les trois modèles: *Flair*, *Spacy*, *CRF* et *BLSTM*
- Chaque modèle renvoie ces outputs.

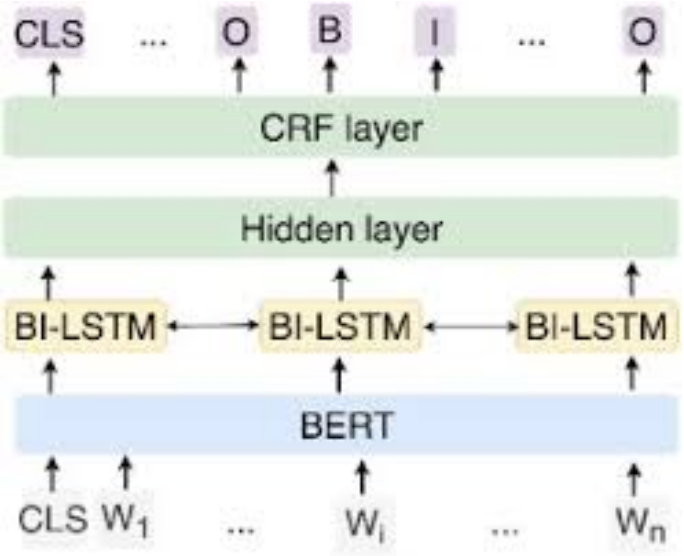


Figure 14: L'architecture Bert-CRF-BLSTM [9]

- L'algorithme sélectionne le modèle qui renvoie le meilleur tag.
- L'algorithme lance un reentraînment en utilisant l'output génère dans l'étape précèdent comme donnée d'entraînement pour les autres modèles.

7

CONCLUSION

Nous avons eu l'occasion de traiter une problématique très répondu dans les domaines de recherche de data science, plus précisément le domaine de NLP dans un cadre BioMedical.
Durant ce projet, nous étions amenées à réaliser un NER sur des différentes thématiques BioMedical.
La richesse de ce domaine, nous a permet de découvrir, utiliser et expérimenter plusieurs architectures d'apprentissage automatique profond.
Nous avons constaté qu'il est toujours possible de combiner plusieurs architectures adaptées pour améliorer les performances des modèles. Une tâche importante d'un data scientist
Nous avons constaté qu'il est toujours possible de combiner plusieurs architectures adaptées pour améliorer les performances des modèles.

References

[1] John Lafferty, Andrew McCallum, Fernando Pereira *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001), 2001

[2] Xuezhe Ma and Eduard Hovy *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF*,
2016

[3] Eliyahu Kiperwasser,Yoav Goldberg *Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations*,2016

[4] Savelie Cornegruta, Robert Bakewell, Samuel Withey and Giovanni Montana,*Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks*,2016

[5] BERT Explained:State of the art language model for NLP, *https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270*

[6] Alan Akbik, Duncan Blythe, Roland Vollgraf *Contextual String Embeddings for Sequence Labeling*

[7] Hidayat Ur Rahman,Thomas Hahn,Dr. Richard Segall*Biomedical Disease Name Entity Recognition Using,NCBI Corpus*

[8] *https://towardsdatascience.com/from-pre-trained-word-embeddings-to-pre-trained-language-models-focus-on-bert-343815627598*

[9] *https://www.sciencedirect.com/science/article/abs/pii/S1386505619310068*

[10] *https://spacy.io/usage/training*