

Reconnaissance des entités médicales

**Réalisé par : EL JAOUHARI Hajar et JAMAL Fedoua**

**Encadré par : RHANOUI Maryem et MIKRAME Mounia**

Table des matières

Page de garde [1](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939653)

[Introduction 3](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939654)

[Problème 3](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939655)

[Observation 3](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939656)

[Approche 3](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939657)

[Explication du modéle 4](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939658)

[Explication de la Data 4](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939659)

[Library 4](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939660)

[Outil utilisé : Google collab 4](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939661)

[Revue de littérature 5](#_Toc93939662)

[Réalisation 8](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939663)

[Le modèle 1 NER\_spacy : 8](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939664)

[Le modèle 2 : CRF 15](#_Toc93939665)

[Le modèle 3 : CNN 24](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939666)

[Comparaison entre accuracy des modèles 25](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939667)

[Conclusion 25](file:///D:\Documents\S5\Ingénierie%20cognitive\Projet\Dataset\rapport%20Ner.docx#_Toc93939668)



Introduction

Aujourd'hui les entreprises de technologie de la santé visent à connecter les communautés médicales avec des millions de patients à travers le pays. Pour permettre aux médecins de répertorier leurs services et de gérer les interactions avec les patients et fournit des services aux patients tels que la réservation, d'interactions avec les médecins et la commande de médicaments en ligne. Les médecins peuvent facilement organiser des rendez-vous, suivre les dossiers médicaux antérieurs et fournir des ordonnances électroniques.

## Problème

Alors il nécessite un modèle prédictif qui peut identifier la maladie et le traitement à partir de l'interaction du patient avec le médecin ou de la commande de médicaments en ligne.

## Observation

* Ici, l'idée est d'extraire les informations utiles du texte et en plus de cela, l'entité des informations utiles doit également être identifiée.
* Parce que les informations que nous prévoyons d'extraire de la déclaration/des interactions entre le médecin et les patients sont liées à des termes médicaux qui ne sont pas des mots d'usage courant.

## Approche

En observant l'exigence, il est clairement visible que nous devons traiter la phrase textuelle et identifier les entités telles que Maladie et Traitement. Nous pouvons prédire toutes ces exigences en utilisant:

* CRF (Conditional Random Field) classifier
* Random Forest Classifier
* HMM (Hidden Markov Model)
* CNN
* Spacy



# Explication du modéle

1. Pour la reconnaissance d'entité médicale, nous construisons un modèle CRF qui utilise un NER personnalisé car les entités que nous prévoyons de reconnaître sont toutes liées à des terminologies médicales.
2. On va utiliser aussi le modèle Ner\_spacy et CNN.
3. Heureusement !! nous avons reçu un ensemble de données qui contient des phrases avec des labels, labels/tags sont très importantes pour la reconnaissance des entités.

# Explication de la Data

Nous avons quatre fichiers de données :

1. Train Sentence Dataset
2. Train Label Dataset
3. Test Sentence Dataset
4. Test Label Dataset

Le fichier de Sentence contient toutes les interactions entre les patients et le médecin et le fichier de Label contient toutes les balises d'entité pour des mots particuliers organisés selon la phrase. Nous avons besoin de faire peu de prétraitement lors de l'accès à l'ensemble de données.

# Library

1. Pandas - Dataframe, Content storage and processing
2. Regular Expression (re) - Identify the textual pattern
3. SpaCy - NLP, POS tag check
4. Warnings - To avoid warning messages
5. Sklearn\_CRFsuite - Model building and Evaluation

# Outil utilisé : Google collab

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l’apprentissage automatique. Cette plateforme permet d’entraîner des modèles de Machine Learning directement dans le cloud. Sans donc avoir besoin d’installer quoi que ce soit sur notre ordinateur à l’exception d’un navigateur.

 Revue de littérature

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Perper name | Few-shot Learning for Named Entity Recognition in Medical Text | Named Entity Recognition with Bidirectional LSTM-CNNs | Biomedical Named Entity Recognition at Scale | a transferable clinical natural language processing model for electronic health records |
| Objectifs | améliorer les modèles NLP  (1) initialisation par couche avec des poids pré-entraînés, (2) réglage des hyperparamètres, (3) combinant des données de pré-entraînement, (4) des intégrations de mots personnalisées et (5) l'optimisation des mots hors vocabulaire (OOV) | détecter automatiquement les caractéristiques au niveau des mots et des caractères à l'aide d'une architecture hybride bidirectionnelle LSTM et CNN, éliminant ainsi le besoin de la plupart des ingénieries de caractéristiques et proposons également une nouvelle méthode d'encodage des correspondances partielles de lexiques dans les réseaux de neurones et la comparons aux approches existantes | la détection d'état d'assertion, la résolution d'entités, l'extraction de relations et la désidentification. | reconnaître sept catégories : noms de médicaments, voie, fréquence, dosage, force, forme, durée. Le modèle a d'abord été pré-entraîné de manière auto-supervisée en prédisant le mot suivant |
| Source et Description de la base de données | **-CoNLL-2003 :** est un ensemble de données de reconnaissance d'entités nommées publié dans le cadre de la tâche partagée CoNLL-2003 Les données se composent de huit fichiers couvrant deux langues : l'anglais et l'allemand. Pour chacune des langues, il existe un fichier de formation, un fichier de développement, un fichier de test et un fichier volumineux avec des données non annotées.  **-III ( MIMIC-III ) :**  L'ensemble de données Medical Information Mart for Intensive Care est une vaste collection de dossiers médicaux anonymisés et accessibles au public. Chaque enregistrement de l'ensemble de données comprend des codes CIM-9, qui identifient les diagnostics et les procédures effectuées. | **-DBpedia** **« base de données » :** est un projet visant à extraire un contenu structuré à partir des informations créées dans le projet Wikipédia.  **-CoNLL-2003**  - **OntoNotes 5.0** : is a large corpus comprising various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows) in three languages (English, Chinese, and Arabic) with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference). | **-BC5CDR** corpus consists of 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases and 3116 chemical-disease interactions.  **-NCBI Disease** corpus consists of 793 PubMed abstracts, which are separated into training (593), development (100) and test (100) subsets. The NCBI Disease corpus is annotated with disease mentions, using concept identifiers from either MeSH or OMIM.  **-JNLPBA** is a biomedical dataset that comes from the GENIA version 3.02 corpus (Kim et al., 2003). It was created with a controlled search on MEDLINE.  **-BC4CHEMD** is a collection of 10,000 PubMed abstracts that contain a total of 84,355 chemical entity mentions labeled manually by expert chemistry literature curators.  **-LINNAEUS** is a general-purpose dictionary matching software, capable of processing multiple types of document formats in the biomedical domain | **-III ( MIMIC-III ) :**  L'ensemble de données Medical Information Mart for Intensive Care est une vaste collection de dossiers médicaux anonymisés et accessibles au public. Chaque enregistrement de l'ensemble de données comprend des codes CIM-9, qui identifient les diagnostics et les procédures effectuées. |
| T echnologies | -Deep neural network models  -NLP  -NER | - [CNN BiLSTM](https://paperswithcode.com/method/cnn-bilstm)  - [BiLSTM](https://paperswithcode.com/method/bilstm)  -[Convolution](https://paperswithcode.com/method/convolution)  • [LSTM](https://paperswithcode.com/method/lstm) • [Sigmoid Activation](https://paperswithcode.com/method/sigmoid-activation) • [Tanh Activation](https://paperswithcode.com/method/tanh-activation) | [Adam](https://paperswithcode.com/method/adam)  • Abandon de l' [attention](https://paperswithcode.com/method/attention-dropout)  • [BERT](https://paperswithcode.com/method/bert)  • [Connexions denses](https://paperswithcode.com/method/dense-connections)  • [Abandon](https://paperswithcode.com/method/dropout)  • [GELU](https://paperswithcode.com/method/gelu)  • [Normalisation](https://paperswithcode.com/method/layer-normalization) de couche • [Échauffement linéaire avec décroissance linéaire](https://paperswithcode.com/method/linear-warmup-with-linear-decay)  • [Attention multi-](https://paperswithcode.com/method/multi-head-attention) têtes • [Connexion résiduelle](https://paperswithcode.com/method/residual-connection)  • [Attention au produit](https://paperswithcode.com/method/scaled) scalé • [Softmax](https://paperswithcode.com/method/softmax)  • [Perte](https://paperswithcode.com/method/weight-decay) de poids • [WordPiece](https://paperswithcode.com/method/wordpiece) | NER |
| Résultats & performance | Experimental results show that the F1 score of 69.3% achievable by state-of-the-art models can be improved to 78.87%. | -F1 (Bi-LSTM-CNN)=91.62  -F1 (Chiu and Nichols (2016))= 86.19 | BLSTM-CNN-Char (SparkNLP)= 89.13  Spark PNL=89,73 | NER= 0,944 |

Réalisation

## Le modèle 1 NER\_spacy :

* Libraries and Packages

Une image contenant texte

Description générée automatiquement

* Common Functions
* Preprocessing Functions

Une image contenant texte

Description générée automatiquement

* Post-processing functions

Une image contenant texte

Description générée automatiquement



* Feature Extraction

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

* Pre-processing

L'ensemble de données fourni se présente sous la forme d'un mot par ligne. Comprenons le format des données ci-dessous :

* Supposons qu'il y ait x mots dans une phrase, alors il y aura x lignes continues avec un mot dans chaque ligne.
* De plus, les deux phrases sont séparées par des lignes vides. Les étiquettes des données suivent le même format.

Nous devons pré-traiter les données pour récupérer les phrases complètes et leurs étiquettes.

Une image contenant table

Description générée automatiquement

L'ensemble de données ci-dessus concerne les étiquettes correspondant aux maladies et au traitement. Trois étiquettes ont été utilisées dans cet ensemble de données : O, D et T, qui correspondent respectivement à Other, \*\*Disease\*\* et \*\*Treatment\*\*, entre chaque deux phrases on a 2 lignes vides et chaque phrase contient au moins une maladie D et un traitement T.



* Extract POS information using SpaCy

Une image contenant texte

Description générée automatiquement

* Frequence of words for POS tag as NOUN & PROPN

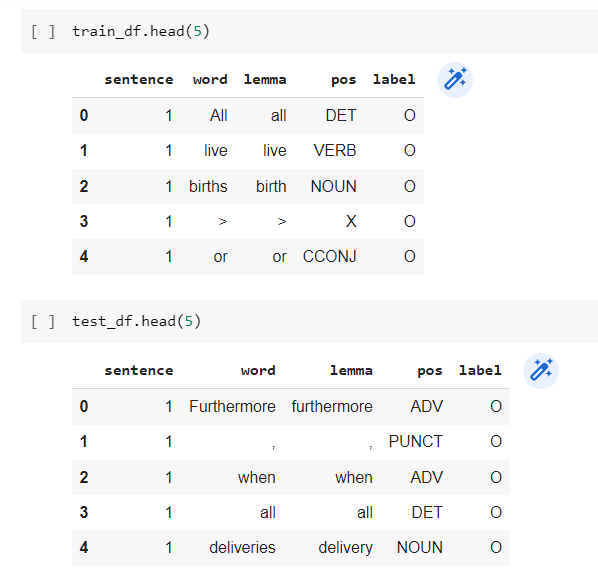
Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

* Dataframe (Sentence, word, POS) visualization



* Sentense-wise detail dataframe preparation

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

* Input and Target Variable

Une image contenant texte

Description générée automatiquement

* Building CRF model using sklearn

Une image contenant texte

Description générée automatiquement

* Evaluate model (F1-score)

Une image contenant texte

Description générée automatiquement

* Predict Disease and Treatment

Une image contenant table

Description générée automatiquement

* Dictionary Preparation

Une image contenant texte

Description générée automatiquement

Une image contenant texte

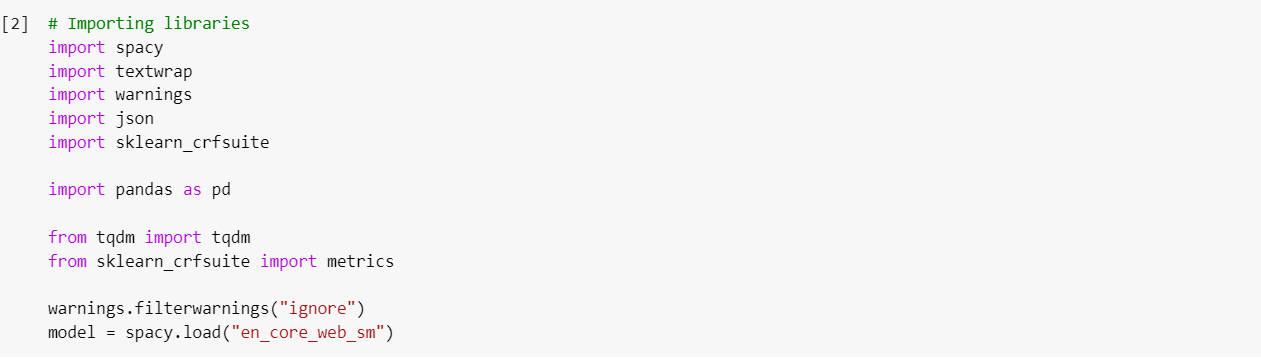
Description générée automatiquement

## Le modèle 2 : CRF

* Libraries and Packages

Une image contenant texte

Description générée automatiquement



* Fonction pour lire le fichier

Une image contenant texte

Description générée automatiquement

* Compter le nombre de sentences(phrases) dans le traitement train et test de la dataset

Une image contenant texte

Description générée automatiquement

* Comptez le nombre de lignes d'étiquettes dans la dataset de train et de test traités.

Une image contenant texte

Description générée automatiquement

* Fonction pour lire le fichier

Une image contenant texte

Description générée automatiquement

* Compter le nombre de sentences(phrases) dans le traitement train et test de la dataset

Une image contenant texte

Description générée automatiquement

* Comptez le nombre de lignes d'étiquettes dans la dataset de train et de test traités.

Une image contenant texte

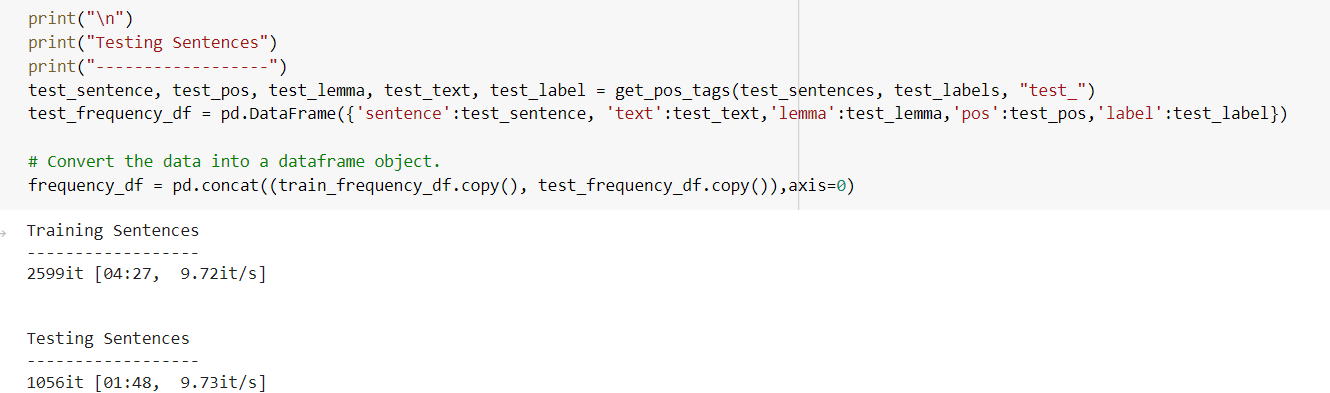
Description générée automatiquement

* Identification des concepts :

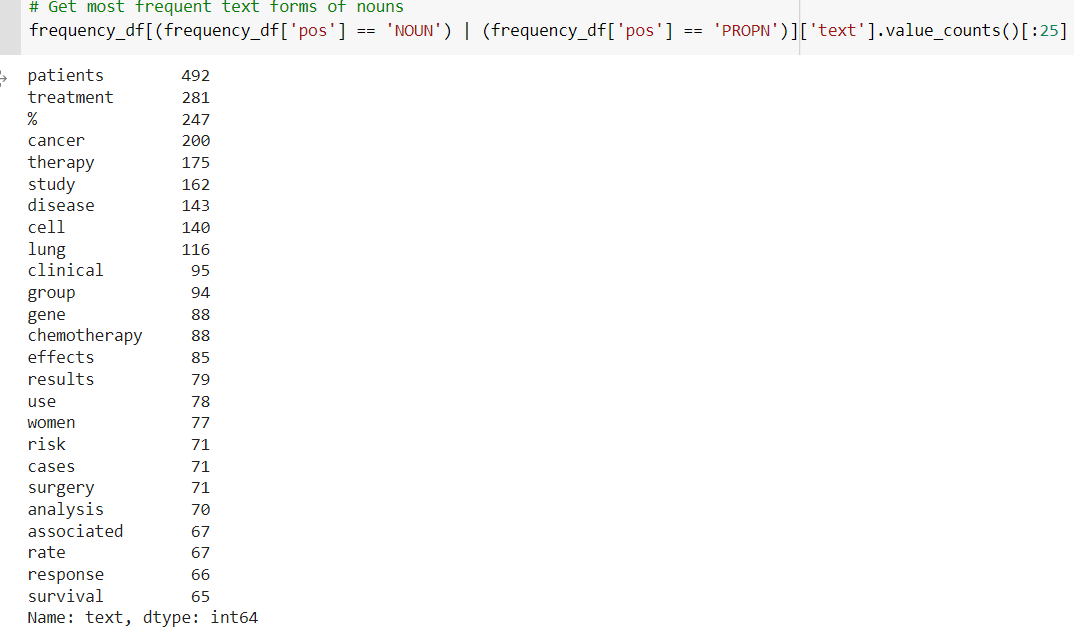
Nous allons d'abord explorer quels sont les différents concepts présents dans l'ensemble de données. Pour cela, nous allons utiliser PoS tagging. Extraire les tokens qui ont NOUN ou PROPN comme étiquette PoS et trouver leur fréquence.

Une image contenant texte

Description générée automatiquement



* Afficher les 25 tokens les plus courants avec NOUN ou PROPN PoS tags.



Une image contenant texte

Description générée automatiquement

* Définition des features pour CRF

Nous avons défini les features suivantes pour la construction du modele CRF :

f1 = le mot d'entrée est en minuscule ;

f2 = les 3 derniers caractères du mot ;

f3 = 2 derniers caractères du mot ;

f4 = 1 ; si le mot est en majuscule ; sinon, 0

f5 = 1 ; si le mot est un nombre ; sinon, 0

f6 = 1 ; si le mot commence par une majuscule ; sinon, 0

f7 = 1 ; si la balise PoS du mot est Noun ou Pronoun ; sinon, 0

f8 = 1 ; si le PoS Tag du mot est Noun ou Pronoun ; sinon, 0

f9 = B ; si début

f10 = E ; si fin

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

* Obtenir les features

Ecrivez un code/fonction pour obtenir les features d'une phrase(sentences).

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

* Écrire un code pour obtenir les features d'une sentence.

Une image contenant texte

Description générée automatiquement

* Définir les variables d'entrée et les variables cibles (input and target)

Calculer correctement les matrices de séquences X et Y pour les données train et test. Vérifier que les phrases(sentences) et les étiquettes sont traitées

Définir les valeurs des features de chaque sentence comme variable d'entrée pour le modèle CRF dans les ensembles de données de test et de train.

Une image contenant texte

Description générée automatiquement

* Définir les labels comme la variable cible (target)pour les données test et train.

Une image contenant texte

Description générée automatiquement

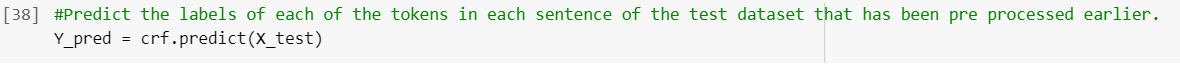
* Construire le model CRF

Une image contenant texte

Description générée automatiquement

* Évaluation

Prédire les labels de chacun des tokens dans chaque phrase de l'ensemble de données de test qui a été pré-traité précédemment.



Calculer le f1 score en utilisant les lables actuelles et les lables Prédites de la dataset test.

Une image contenant texte

Description générée automatiquement

* Identification des maladies et des traitements en utulisant NER :

Nous utilisons maintenant la prédiction du modèle CRF pour préparer un registre des maladies identifiées dans le corpus et des traitements utilisés pour ces maladies.

Créer la logique pour obtenir tous les labels de traitements prédits (T) correspondant à chaque label de maladie (D) dans la dataset test.

Une image contenant table

Description générée automatiquement

Une image contenant table

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

* Prédire le traitement de la maladie dont le nom est : 'hereditary retinoblastoma'.

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

* Mesure de la performance de notre modéle

Une image contenant table

Description générée automatiquement

* Prédire le traitement de la maladie dont le nom est : 'hereditary retinoblastoma'.

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

* Mesure de la performance de notre modéle

Une image contenant table

Description générée automatiquement

## Le modèle 3 : CNN

Dans ce modèle on a regroupé les fichiers .txt en un seul fichier qui s’appelle Train\_and\_test.txt.



CNN :1D

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Une image contenant table

Description générée automatiquement

## Comparaison entre accuracy des modèles

|  |  |
| --- | --- |
| Modèle | Accuracy |
| NER\_spacy | 91.7% |
| CRF | 92.01% |
| CNN | 8% |

## 

D’après l’exécution des différents modèles, on déduit que nous devons déployer le deuxième modèle, car tout simplement c’est celui qui nous a donné de très bonne performance au niveau de la prédiction d’une valeur égale 92.01% au lieu de celle du premier modèle qui égale 91.7% et le troisième modèle qui égale 8%.

Conclusion

Ce projet présente une approche médicale NER qui intègre un dictionnaire de domaine médical et des modèles de langage préformés dans la reconnaissance des entités médicales nommées. En nous concentrant sur la façon d'extraire des connaissances à partir de textes médicaux étiquetés.

La reconnaissance d'entités médicales nommées à l'aide de textes et de dossiers médicaux étiquetés est une tâche difficile. Ce travail crée un moyen d'utiliser un dictionnaire médical, des modèles préformés dans le domaine de NLP et plus précisément le Named Entity Recognition pour extraire des textes médicaux étiquetés.